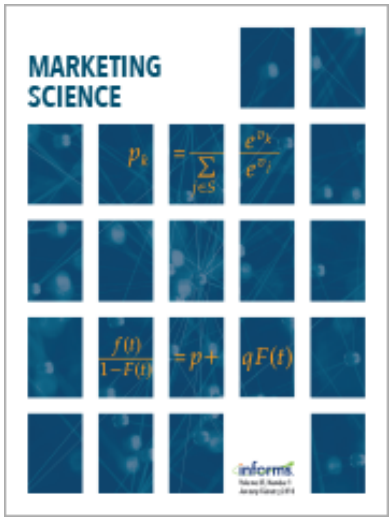


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Privacy-Preserving Data Fusion

Longxiu Tian, Dana Turjeman, Samuel Levy

To cite this article:
Longxiu Tian, Dana Turjeman, Samuel Levy (2026) Privacy-Preserving Data Fusion. Marketing Science

Published online in Articles in Advance 01 Apr 2026

. https://doi.org/10.1287/mksc.2023.0068

https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–24
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Privacy-Preserving Data Fusion

Longxiu Tian,[a,*] Dana Turjeman,[b] Samuel Levy[c]

[a] Kenan–Flagler Business School, University of North Carolina, Chapel Hill, North Carolina 27599; [b] Arison School of Business, Reichman University, Herzliya 46101, Israel; [c] Darden School of Business, University of Virginia, Charlottesville, Virginia 22903

*Corresponding author

Contact: longxiu@unc.edu, https://orcid.org/0000-0001-6257-7583 (LT); dana.turjeman@runi.ac.il, https://orcid.org/0000-0003-1445-2983 (DT); levys@darden.virginia.edu, https://orcid.org/0000-0002-1188-648X (SL)

Received: February 14, 2023
Revised: May 22, 2024; May 12, 2025; October 31, 2025
Accepted: November 10, 2025
Published Online in Articles in Advance: April 1, 2026

https://doi.org/10.1287/mksc.2023.0068

Abstract. Data fusion combines multiple data sets to improve inference, but it can inadvertently reveal user identities, even when data appear anonymous. We introduce a privacy-preserving data fusion (PPDF) methodology, combining an expressive multidata set variational autoencoder architecture and a differentially private training procedure. Additionally, to aid managerial interpretability, we develop a posterior reidentification metric to quantify individual-level reidentification risks under data fusion and formally derive the analytical bound to the heightened privacy risks posed by data fusion. We showcase PPDF's abilities by fusing an anonymous customer satisfaction survey and a customer relationship management database of a large U.S. telecom provider. In a predictive churn-prevention campaign, PPDF achieves a 1.46% campaign lift without revealing any user identities compared with a 1.66% lift in a model without privacy guarantees, in which nearly 7% of users are reidentified. A heterogeneity analysis further reveals these individuals as outlier customers who exhibit persistently higher risk of reidentification, underscoring the nuanced trade-off between individual-level privacy guarantees and overall inference quality. More broadly, PPDF enables organizations to harness the full power of data fusion without violating user privacy—offering a practical solution for firms, governments, and researchers seeking to extract value from sensitive data while ensuring individual confidentiality.

History: Olivier Toubia served as the senior editor.

Funding: This research was supported by the Israel Science Foundation (ISF) [Grant 287/23].

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mksc.2023.0068.

Keywords: differential privacy • data fusion • variational autoencoders • generative modeling • selection bias

1. Introduction

In today's world, privacy stands both as a legal imperative and as a core human right, placing firms' marketing activities as a focal point of increasing scrutiny. Regulations, such as the General Data Protection Regulation and various state and national mandates, have broadened the scope of what constitutes personally identifiable information (PII), extending its scope to cover attributes that even indirectly identify customers. This broader definition of PII necessitates a reassessment of data governance and usage practices, particularly regarding data linkage techniques, which pose the unique risk of *reidentifying* individuals (Lobschat et al. 2021, Hassani et al. 2024).[1] Data linkage or data fusion is a foundational technique in marketing aimed at merging data sets to improve analytics, which must now adhere to stricter privacy standards and expectations.

Take, for example, a company that wants to boost customer satisfaction and loyalty within its own customer base. It may do so by collecting *anonymous* customer satisfaction responses either through external surveyors or independently in order to capture genuine feedback from its customers (Bradburn et al. 1979). By fusing such anonymized attitudinal survey data with behavioral insights from their customer relationship management (CRM) system, the company may uncover key satisfaction drivers crucial for its growth (Malshe et al. 2020).

In practice, such fusions are typically carried out either deterministically (e.g., by directly matching PII if it is collected) or probabilistically (e.g., via data fusion algorithms leveraging common variables). Privacy safeguards

in these processes, if used at all, are usually minimal. The most common practice is to undertake a deidentification step (i.e., post hoc removal of PII). As a result, managers face a critical trade-off; despite its widespread applications and benefits, data fusion introduces significant and unique privacy challenges for customer-level data.[2] Notably, Sweeney (1997), Narayanan and Shmatikov (2008), and Li et al. (2023) show that data fusion on nominally anonymized data can lead to "linkage attacks," where the combination of data sets exposes sensitive and identifiable information of individuals. The core motivation of our research is that even with anonymous or deidentified data, the fusion process itself may facilitate reidentification by inadvertently linking an individual's records across data sets, thus revealing their sensitive information.[3] Hence, fusing data sets, such as anonymous survey responses, with CRM data poses risks to various stakeholders. Customers whose data are fused stand as the primary at-risk group followed by the companies conducting and commissioning the surveys. Breaches that compromise customer and respondent anonymity can erode trust and subject these companies to legal repercussions, fines, and reputational damage (Ruohonen and Hjerppe 2022, Turjeman and Feinberg 2024), underscoring the difficulty of aligning in-depth analytics with the imperative of maintaining customer privacy. Furthermore, beyond the inherent privacy risk itself, to date, objective quantification of reidentification risks is still limited.

Prior work has examined several approaches that firms use in practice to safeguard privacy in data sharing, including anonymization, deidentification, aggregation, and contractual restrictions on data use. Although these methods offer *some* level of protection, research consistently shows that they do not eliminate the risk of reidentification once data sets are linked. Even when firms comply with legal requirements, such as the General Data Protection Regulation in the European Union, the residual reidentification risk persists and can expose both customers and firms to harm. This highlights a persistent managerial dilemma; improving fusion quality enhances analytic and predictive performance, but doing so without formal safeguards increases privacy risks, resulting in legal, reputational, and material consequences.

This research introduces privacy-preserving data fusion (PPDF), a novel framework enabling firms to fuse disparate customer-level data sets while providing quantifiable and tunable guarantees that customers are not reidentified in the fusion process. PPDF's objective function goes beyond accuracy, taking into account the likelihood of reidentifying individuals in the data sets. This is crucial for meeting the increasing demands for compliance and efficiency in data-driven industries (Hassani et al. 2024). To realize a coherent data fusion methodology that ensures the utility of the learned data distributions while maintaining rigorous privacy standards, PPDF is built from first principles based on differential privacy (DP) (Dwork et al. 2006b), variational autoencoders (VAEs) (Kingma and Welling 2019), and normalizing flows (Rezende and Mohamed 2015). PPDF stands out as a fully generative data fusion technique; it fuses data sets without the necessity of predefining a discriminative model, and crucially, any downstream inference also inherits its privacy guarantees.

Our empirical applications showcase the practical value of PPDF via a churn-prevention exercise that leverages data fusion. Under a modest privacy budget[4] ($\varepsilon = 5.0$), we attain a predicted campaign lift of 1.46% while entirely eliminating reidentification of customers; this is only 0.2 percentage points shy of the best nonprivacy-preserving benchmark at 1.66% that albeit being slightly more accurate, resulted in nearly 7% of customers being reidentified. These findings suggest that firms can, in practice, ensure not only the preservation of customer-level privacy but also, maintain the requisite degree of accuracy for their data fusion workflow. Additionally, a heterogeneity analysis of users otherwise reidentified without PPDF reveals that there is a persistent segment of "edge case" customers who consistently face the highest privacy threats under conventional data fusion methods. These findings demonstrate that firms seeking to enhance retention can feasibly fuse survey-based attitudinal insights with CRM data, all without compromising individual privacy or incurring large losses in campaign performance.

Conceptually, this research shows that data fusion imposes privacy risks above and beyond those associated with individual data sets, thus motivating a holistic approach to privacy preservation in multisource marketing analytics. Methodologically, our proposed PPDF framework addresses these fusion-specific vulnerabilities through a generative model that integrates variational autoencoders, normalizing flows, and differential privacy. To aid in the interpretability and actionability of PPDF, this research introduces a posterior reidentification metric that at the individual level, quantifies how many users' records are indistinguishable from their own crossimputed data—thereby providing a transparent measure of reidentification risk under data fusion. As noted by Klymenko et al. (2024), one of the main challenges in the adoption of privacy-enhancing technologies is the lack of clear privacy metrics. Our framework aims to bridge this gap by offering a quantifiable, actionable metric that accompanies their data fusion practices, enabling firms to benchmark and set informed privacy policies while maintaining data utility. The tandem of this metric and PPDF ensures that firms can harness the enriched insights afforded by fusing disparate data sets while systematically mitigating user-level reidentification risks.

The remainder of the paper is structured as follows. In Section 2, we review the literature on data fusion and

privacy-preserving methods. In Section 3, we describe our data sources and institutional contexts. We then derive how multisource data fusion introduces additional privacy risks over and above single-source analysis in Section 4. In Section 5, we present our proposed PPDF framework and accompanying posterior reidentification metric. We report empirical findings in Section 6, including an in-depth heterogeneity analysis exploring the potential process underlying reidentification risk. We discuss limitations and assumptions in Appendix A. Finally, Appendix B concludes, outlining both practical implications and potential future research directions.

2. Related Literature

Data fusion and record linkage, originally discussed by Dunn (1946) and widely adopted by leading technology firms, combine multiple data sources to make inferences that surpass the accuracy, generalizability, and utility achievable with any single data set alone. Data fusion has been shown to effectively enhance inference across various domains (Bradlow and Zaslavsky 1999; Swait and Andrews 2003; Feit et al. 2010, 2013; Unger et al. 2018; McCarthy and Oblander 2021).

Gilula et al. (2006) treat data fusion as a problem of learning the joint distribution of two disjoint sets of variables when for any given respondent, only one set is observed. As an illustration, let age denote the variable that appears in both surveys, income in one data set, and purchase history in the other. By assuming conditional independence—that is, $P(\text{income}, \text{purchase}|\text{age}) = P(\text{income}|\text{age}) \cdot P(\text{purchase}|\text{age})$—the task of "crossimputing" the missing block (income or purchase history) becomes an ordinary missing data exercise. Practically, the firm can then draw from the predictive (posterior) distribution of the complete data conditional on the observed variables for each record.

Although data fusion has been proven to be useful in many contexts, when fusing individual-level data sets, common variables manifest as *quasi-identifiers* as their commonality between deidentified component data sets and those that do contain PII can serve as a mechanism to reidentify individuals (Sweeney 1997, Narayanan and Shmatikov 2008). Although firms may be tempted to prevent crossdata sets matching altogether, such fragmentation can impede personalization and inadvertently harm customers and firms (Lin and Misra 2022, Korganbekova 2023, Neumann et al. 2024). Table 1 summarizes representative work on data fusion and privacy. Of note, prior marketing fusion efforts that have relied on aggregate-disaggregate merges (e.g., Feit et al. 2013, McCarthy and Oblander 2021) inherently present lower privacy risk than individual-level data fusion.

PPDF's formal privacy guarantee, referred to as the "privacy budget," is operationalized through its *differentially private* training. Differential privacy has emerged as the dominant paradigm for privacy protection in computer science because it provides a mathematically rigorous and generalizable guarantee that holds regardless of the attacker's auxiliary information (Dwork et al. 2006a). The core intuition is simple; by adding carefully calibrated random noise during training, DP ensures that the inclusion or exclusion of any single individual's data has limited effect on the corresponding outputs, thereby reducing the risk that an individual will be assumed to be included in the data set.

This formal guarantee makes DP qualitatively stronger than deterministic approaches, such as anonymization or deidentification, which can often be reversed through linkage attacks. Alternative privacy-preserving technologies, like the classic k- and l-anonymity (Sweeney 2002), and more modern solutions—homomorphic encryption (Dehghan et al. 2022), deep learning (Takagi et al. 2020, Anand and Lee 2023), federated learning (Huang et al. 2023), and *local* differential privacy (Kasiviswanathan et al. 2011)—address individual-level fusion or data sharing, but they often have drawbacks in scalability or generalizability (Shelake and Shekokar 2017, Ding et al. 2019, Gati et al. 2021, Hassani et al. 2024). Moreover, to the best of our knowledge, none of the current privacy-preserving technologies provide ex post reidentification risk metrics (Carey et al. 2023), and most are domain specific (Kaissis et al. 2021, Wang and Yang 2024). Beyond its widespread adoption in computer science and policy contexts (e.g., U.S. Census Bureau 2021), DP is beginning to see applications beyond computer science, including in pricing and marketing analytics (e.g., Chen et al. 2022), in platform-level marketing systems (e.g., Rogers et al. 2021), and in data market design contexts (e.g., Liu et al. 2021). By situating our work within this emerging stream, we highlight the novelty of adapting DP methods for privacy-preserving data fusion in marketing contexts.

Our proposed framework extends the literature by first building on the established computer science finding that data fusion magnifies reidentification risks through linkage attacks and formalizing how these risks manifest in marketing-relevant contexts, such as CRM-survey fusion. Second, we propose a scalable and expressive probabilistic data fusion methodology with tunable privacy guarantees (Kim and Tanuwidjaja 2021, Evans et al. 2023). Finally, we propose a theoretically grounded and managerially interpretable reidentification risk metric curated for data fusion tasks.

3. Data Description

We provide an empirical illustration based on an actual managerial challenge presented to us by a leading U.S. telecom carrier. The data available from the telecom carrier include three data sets.

Table 1. Representative Literature on Data Fusion and Privacy and the Positioning of Our Privacy-Preserving Data Fusion Approach

Paper	Focus/domain	Method/key contribution
Dunn (1946)	Early record linkage	Combined multiple population data sets for improved accuracy; foundational work on data fusion
Sweeney (1997)	Privacy attacks via linkage	Demonstrated how quasi-identifiers (e.g., demographics) allow reidentification of supposedly "anonymous" health data
Sweeney (2002)	*k*-anonymity for privacy	Introduced *k*-anonymity as an approach to obscure individuals among at least *k* others; limited scalability for high-dimensional data
Bradlow and Zaslavsky (1999)	Missing data in surveys	Applied data fusion to handle incomplete survey responses, showing how linked data can enrich inferences
Narayanan and Shmatikov (2008)	Reidentification from public data	Matched the Netflix data set with IMDb profiles; highlighted privacy risks when deidentified data are combined
Feit et al. (2010, 2013)	Marketing data fusion	Fused disaggregate choice data with CRM or media platforms for enriched preference estimation; did not address individual privacy
Shelake and Shekokar (2017)	Privacy-preserving data integration	Reviewed cryptographic, anonymization, noise-based methods; emphasized reidentification vulnerabilities in multisource data linkage
McCarthy and Oblander (2021)	Aggregate-disaggregate fusion	Combined aggregated and disaggregated data for scalable customer base analysis; limited privacy risk because most data were aggregate
Lin and Misra (2022), Korganbekova (2023), and Neumann et al. (2024)	Consequences of *not* fusing	Showed how privacy-driven prevention of identity matching causes "data deserts," fragmenting insights and harming targeting or small players
Gati et al. (2021)	Differentially private fusion for cyber-physical-social systems	Surveyed privacy techniques (DP, encryption) for Internet of Things (IoT) data fusion; stressed the need for robust privacy frameworks at scale
Cai et al. (2022)	Multimodal data fusion	Used deep learning to handle temporal alignment and synchronization at the variable level; employs differential privacy but requires variable-level synchronization
Dehghan et al. (2022)	Homomorphic encryption	Proposed a privacy-preserving data fusion protocol based on homomorphic encryption; secure but high computational overhead
Wang and Yang (2024)	Vertical federated learning for traffic data	Proposed a privacy-preserving data fusion approach in transportation reliant on domain-specific physics constraints; not easily generalized
Anand and Lee (2023)	Deep learning for data sharing	Developed a privacy-friendly approach to transfer data representations without exposing raw data; not specifically for generative fusion
Carey et al. (2023)	Measuring reidentification risk	Formal framework quantifying attacker success probabilities postrelease; no integrated private training phase
Our paper: PPDF	Privacy-preserving *generative* fusion in marketing	Proposes a variational autoencoder + DP framework for fusing multiple data sets *without* revealing identities; explicitly quantifies individual reidentification risk *and* ensures robust model accuracy

1. One data set includes responses from an industry-standard, anonymous customer satisfaction survey, hereafter called the "external survey." The survey was carried out by an external surveying company that continuously conducts market research on customer satisfaction. Survey outcomes include measures of satisfaction with different service components as well as an overall likelihood to recommend (LTR). The resulting measures of the external survey are a well-known industry standard used by customers and managers to compare quality of service. We provide a set of questions from the external survey in Online Appendix EC.5.

2. One data set is responses of an additional internal validation survey conducted by the telecom carrier. This internal survey is identical to the external survey except that it provides a ground-truth mechanism; each recipient received a unique email link, enabling the association of survey responses with customer identifications.

3. One data set includes detailed CRM data of customers randomly selected from the carrier's full customer base (3.6 million accounts, which are roughly 2.0%–5.0% of the full customer base).[5]

The telecom carrier's goal in this workflow is to explore the relationship between customer satisfaction and the likelihood of churning out of the company's services. Customer surveys, although extensive and including many engagement variables, do not include measures of financial outcomes, technical metrics on device and network usage, or plan and service details. Combining the survey data along with the CRM data may enable the carrier to conduct a churn-prevention

campaign and to retain those customers who have a higher likelihood of churning.

Notably, the external survey data should remain anonymous. Therefore, to avoid the risk of reidentification, up to the development of PPDF, the telecom carrier could not have fused the two data sets as they lacked certainty on privacy assurances. Here, the internal validation survey represents a unique, one-time mechanism for ground-truth validation of PPDF. However, the added step of the validation survey is not necessary for future end users of PPDF, and it is presented in this research to provide mechanistic interpretability for the proposed method and allow for a heterogeneity analysis of user-level reidentification risks.

3.1. Internal and External Surveys

Both the internal validation and external surveys were conducted in the third quarter of 2021. In the internal validation survey, an email was sent to 2.1 million of the telecom carrier's customers (randomly selected from the entire customer base except for prepaid customers, who were excluded from all data sets). Emails were sent gradually throughout the surveying period. A total of 20,233 customers responded to the internal survey, which corresponds to a response rate of 0.95%. Participants did not receive compensation for completing the surveys. The external survey was conducted by an external surveying company that continuously conducts market research on customer satisfaction on behalf of market leaders in the wireless telecom sector. Approximately 8,000 respondents responded to the external survey, all of whom were customers of the telecom carrier at the time of responding.

Both surveys were identical and included customer satisfaction/perception questions, questions about engagement with the carrier's products and services, questions about the customer relationship with the telecom carrier, and sociodemographic questions. Participants who did not complete the entire survey were excluded; however, they comprised less than half of a percent of the total sample. The survey taxonomy is composed of four mutually exclusive question types: *identifiers* (e.g., sociodemographics), *relationship* (e.g., plan choice, account type, and devices), *engagements* (e.g., times that a retail store was visited, plan switching, and customer service calls), and *perceptions* (e.g., satisfaction with services and devices and LTR).

3.2. CRM Data

The telecom carrier has extensive and detailed user data from the moment that a user joins. The data available for this project are of two types: common variables and uncommon variables. Common variables are shared between the surveys and the CRM data so that a common variable for a given customer will have the same value or attribute across both data sets.[6] Common variables cover engagement metrics, such as the purchase of a new phone or connected device, reward redemption behavior, and recent visits to a retail location. Uncommon variables are idiosyncratic to their respective data sets. Uncommon variables include the number of lines for the account, the tenure of the customer, the status of the contract, and billing information. The full list of common and uncommon variables is available in Online Appendix EC.5.

4. Data Fusion's Unique Privacy Risks

The focal managerial challenge that this research seeks to address is that in fusing multiple data sets, particularly at the individual (disaggregate) level, the overall risk of reidentification can be substantially higher than the sum of the risks innate to each component data set. Consequently, there is a need for a methodology that explicitly mitigates, controls, and quantifies this reidentification risk while enabling meaningful individual-level data fusion.

In this section, we first provide an illustration to intuit the mechanisms leading to an expansion of privacy risks in a typical data fusion workflow informed by the current practices of the telecom carrier from our empirical example. We then formally define and derive the privacy budget under data fusion (Dwork et al. 2006b), which we analytically show to exceed the sum of the individual data sets' privacy budgets. This serves as the theoretical basis for PPDF's differentially private training, downstream guarantees, and generalizability as detailed in Section 5.

4.1. Illustration and Conceptualization

A seminal work in data privacy, the main finding of Sweeney (1997) is that anonymity during the data collection phase does not guarantee privacy under data fusion. In other words, anonymity is not a synonym for privacy. Consider a company that conducts an anonymous customer satisfaction survey to gain insights into its products and services. Being anonymous, the survey avoids collecting PII, such as names, email addresses, or customer identifications. This design serves two purposes: to encourage candid responses and to assure the respondents of their privacy. Yet, this nominal assurance of privacy often fails in practice because of the existence of variables known as *quasi-identifiers*. Although not conventionally PII as they do not deterministically identify users, quasi-identifiers can enable reidentification when their unique combinations "triangulate" a customer (Dankar et al. 2012) (e.g., answers to demographic questions together with those on how the customer has engaged with the firm).

Consider the following customer profile as indicated by their responses to the anonymous survey: over 90 years old living in zip code 90210 (demographics) and

purchases a new phone every 30 days (engagement). Although the combination of these survey responses likely represents an outlier in the customer base, the customer cannot be said to be reidentified when considering the survey in isolation. However, in eliciting responses on engagements and demographic information, this opens up the possibility of crossreferencing (i.e., fusing) the survey with the firm's CRM, where many of these variables also exist. Our respondent's outlierness in these *common variables* implies a nonnegligible chance of pinpointing the correct customer profile. More troubling, as some form of PII necessarily exists in a firm's CRM database, the customer's name, address, and other PII can now be linked to the survey responses.

When data sets are fused, common variables between the data sets become quasi-identifiers. It is difficult, if not impossible, to imagine common variables in realistic data fusion scenarios that do not serve as quasi-identifiers under a suitably expressive data fusion method. Should at least one of the component data sets contain PII, then the data fusion exercise necessarily results in greater reidentification risk. That is, component data sets that otherwise were anonymized or deidentified ex ante could be reidentified through data fusion with data sets that contain PII because of the existence of common variables serving as quasi-identifiers.

For managers, a linkage attack (Sweeney 2000) and a data fusion task are indistinguishable in terms of implementation and differ only in terms of intent and consent. For instance, a malicious actor could perform a linkage attack by matching unique combinations of demographic and engagement details between an anonymized survey and the firm's CRM database. Conversely, a marketing team could use the same process to gain insights to improve their strategies. Although the intent differs, the data fusion process employs the same quasi-identifiers, leading to similar privacy risks. Therefore, even well-intentioned data fusion tasks can inadvertently increase reidentification risk. The possibility of privacy breaches persists because the methodologies that provide valuable business insights also enable linkage attacks. In essence, replacing the malicious actor with a well-meaning marketer does not eliminate the inherent privacy risks associated with data fusion using customer-level data.

4.2. Privacy Budget Under Data Fusion

Building on the above illustration's intuition that data fusion results in greater reidentification risk, we make use of differential privacy's analytical properties (Dwork et al. 2006a, Steinke 2022) to formally derive the privacy budget of data fusion.

To do so, we first formally define the concepts of quasi-identifiers and privacy budget. For readers who are less familiar, we provide a detailed review of DP's analytical properties in Online Appendix EC.2.1. Let data sets D_1 (e.g., survey without PII) and D_2 (e.g., CRM with PII) share a set of common variables $\mathbf{X}^{(c)}$. A specific common variable, $\mathbf{x}^{(c)}$, is a *quasi-identifier* if it is sufficiently specific to narrow a record (or a small set of records) when data sets are linked on $\mathbf{X}^{(c)}$. Suppose there exists $\mathbf{x}^{(c)}$ such that $\text{count}_{D_1}(\mathbf{x}^{(c)}) \geq 1$ and $\text{count}_{D_2}(\mathbf{x}^{(c)}) = 1$; then, fusing D_1 and D_2 on $\mathbf{X}^{(c)}$ uniquely reidentifies the corresponding individual. More generally, if $\text{count}_{D_2}(\mathbf{x}^{(c)}) \leq k$, the reidentification probability is at least $1/k$. Linking on $\mathbf{X}^{(c)}$ maps survey rows with $\mathbf{x}^{(c)}$ to exactly one (or at most k) CRM record containing PII, yielding unique (or k-set) reidentification. Building on these definitions, the total *privacy budget* (i.e., reidentification risk) of data fusion is greater than the sum of those of its component data sets (Lemmas 1 and 2).

Lemma 1. *A generative model of a single data set D with known linkage between its column-wise partitions has a privacy budget equal to the sum of the privacy budgets of the partitions.*

Lemma 2. *A J-data set fusion using common variables leads to an expansion of the privacy budget by $J\varepsilon$ over and above the sum of the individual data sets' budgets.*

See Online Appendix EC.2.2 for detailed proofs of these lemmas. As a sketch of the formal proofs, Lemma 1 states that under data fusion, a data set can be partitioned into uncommon and common variables, where the latter is to be used to learn the joint distribution across data sets. If a component data set j was considered only in isolation, then a single model can be trained on both data partitions, effectively treating them as one in learning j's data-generating process, where $j = 1, \ldots, J$. Under DP, a single model implies a single privacy budget denoted by ε. Yet, data fusion entails more than just learning J data sets in isolation; in addition to learning each data set's data-generating process, there must also be separate mechanisms to learn the joint distribution across data sets. Under DP, the mechanisms for crossdata set learning are treated as additional models, each with its own privacy budget of ε. Moreover, privacy budgets are additive (Steinke 2022). To this end, we derive (Lemma 2) the total privacy budget under data fusion to be greater than that of its component data sets' sum: specifically, by an amount of $J\varepsilon$. This is precisely because of the existence of J crossdata set learning mechanisms under data fusion that do not exist when modeling each component data set in isolation.

Conceptually, as we move to developing the proposed PPDF methodology in the next section, our goal is to minimize data fusion's privacy budget while balancing fusion accuracy in empirical applications. Whereas this section made use of DP's analytical properties to define data fusion's privacy budget, we next leverage DP's

computational properties (Abadi et al. 2016) to incorporate a tunable privacy budget into a scalable and accurate data fusion framework.

5. Methodology

To motivate and develop PPDF's methodology,[7] we begin below with a simpler case in which we conceptualize partitioning a single data set, in line with Lemma 1 from above. Then, having established how the model works under such oracle conditions, we advance to the focal scenario of combining separate data sources in Section 5.2, where we use a multiencoder and multidecoder architecture to form a joint "full data" distribution (Gilula et al. 2006) and describe its corresponding differentially private training algorithm (Abadi et al. 2016). Finally, we develop a posterior reidentification metric to assess individual-level privacy risks generalizable to any data fusion context in Section 5.3.

5.1. Illustrative Scenario: Single Data Set

We first describe a simpler "oracle" scenario in which a single data set contains *all* variables. Although this scenario does not hold in many practical settings, it serves two key purposes. First, it illustrates the fundamental operation of a variational autoencoder when *no* data fusion is required, showing how latent representations are learned and used for reconstruction in an ideal scenario of a complete data set. Second, it demonstrates how differentially private stochastic gradient descent (DP-SGD) can be seamlessly integrated into VAE training to safeguard individual records, even when all variables are centrally available.

5.1.1. Variational Autoencoder Overview. Figure 1 shows the general architecture of a variational autoencoder trained on a single (oracle) data set. For simplicity, assume that the data are *partitioned* into three segments,

$$\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}_c, \mathbf{x}^{(2)}),$$

where $\mathbf{x}^{(1)}$ denotes variables unique to one data source (e.g., CRM), $\mathbf{x}_c$ denotes variables common across multiple sources (e.g., shared demographics), and $\mathbf{x}^{(2)}$ denotes variables unique to another source (e.g., survey responses). In this illustrative scenario, the analyst has access to *all* of these variables in a single place; no data are truly "missing."

- **Single encoder (inference model).** A neural network encoder $q_\phi(\mathbf{z}|\mathbf{x})$ takes in all parts of $\mathbf{x}$ and outputs a latent representation $\mathbf{z}$; this is also known as a *recognition* or *inference* model.
- **Single decoder (generative model).** A neural network decoder $p_\theta(\tilde{\mathbf{x}}|\mathbf{z})$ reconstructs the original input $\mathbf{x}$ (i.e., $\tilde{\mathbf{x}} \approx \mathbf{x}$). In practice, implemented as multiple subnetworks, each is dedicated to reconstructing a different part of the data set $(\mathbf{x}^{(1)}, \mathbf{x}_c, \mathbf{x}^{(2)})$.

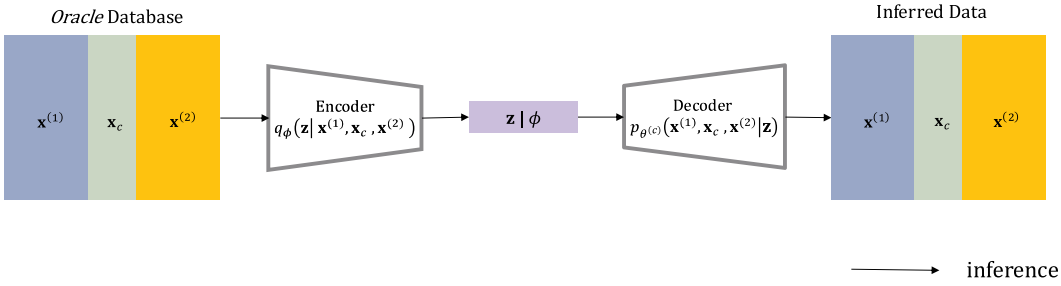
At training time, the VAE attempts to *self-supervise* by comparing $\mathbf{x}$ with $\tilde{\mathbf{x}}$ (i.e., measuring a *reconstruction error*) while simultaneously regularizing $\mathbf{z}$ so that it remains close to a chosen prior distribution (usually, a standard Normal distribution $N(0,1)$). Formally, this can be expressed via the *evidence lower-bound* (ELBO) objective:

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})\big). \quad (1)$$

Maximizing this ELBO is equivalent to maximizing the log-likelihood $\log p_\theta(\mathbf{x})$ up to a constant lower bound, where KL(·) is the Kullback–Leibler divergence penalizing deviations of $q_\phi(\mathbf{z}|\mathbf{x})$ from the prior $p_\theta(\mathbf{z})$. In this oracle setting, a VAE captures intricate relationships among all observed variables in one pass. Although simplified and stylized, this example clarifies the essential mechanics.

1. We feed all data $\mathbf{x}$ into an encoder.
2. The encoder produces latent variables $\mathbf{z}$.
3. A decoder uses $\mathbf{z}$ to reconstruct $\mathbf{x}$.
4. Training refines the encoder and decoder by minimizing reconstruction loss and KL divergence.

Figure 1. Illustration of a Variational Autoencoder of a Single Data Set



Notes. All parts of this full data set and variables $\mathbf{x}^{(1)}$, $\mathbf{x}_c$, and $\mathbf{x}^{(2)}$ are encoded to the latent variables $\mathbf{z}$ through the function $q(.)$ parameterized by the variational vector of parameters ϕ through a neural network. The latent vector $\mathbf{z}$ is then decoded via a decoder parameterized by $\boldsymbol{\theta}$ through another neural network to reconstruct the original data. The VAE is a stochastic computational graph that simultaneously optimizes the variational parameters ϕ and the model parameters $\boldsymbol{\theta}$. This schematic depicts the VAE module of a single data set only; data fusion and privacy protection are introduced later in Figure 3 and Figure 4, respectively.

5.1.2. Where Privacy Comes in. Training the VAE with standard stochastic gradient descent exposes potential privacy risks because the gradient updates can inadvertently reveal sensitive information about individual records in the data set. To mitigate this, we incorporate *differentially private* stochastic gradient descent (Abadi et al. 2016), ensuring that no single observation's unique traits are disclosed through the learning process.

Figure 2 provides a geometric illustration of DP-SGD in a simple two-dimensional latent space. The concentric black ellipses in Figure 2 represent contours of the VAE's loss function; points closer to the center in Figure 2 have lower loss. In a typical (nonprivate) setting, each minibatch of samples produces a gradient (shown by a blue arrow) pointing toward a local descent direction. However, with DP-SGD, each gradient vector is subject to two key modifications.

1. **Gradient clipping.** First, we *clip* each individual gradient so that its Euclidean norm does not exceed a threshold C. Graphically, this corresponds to restricting the gradient to lie *within* the green dashed circles in Figure 2. This bounding step prevents any single record from disproportionately steering the model parameters. Formally, for a minibatch of samples $\{x_1, \ldots, x_N\}$, compute each sample's gradient $g_t(x_i)$ for all $i \in \{1, .., N\}$. Rescale any gradient exceeding norm C down to C, thus ensuring

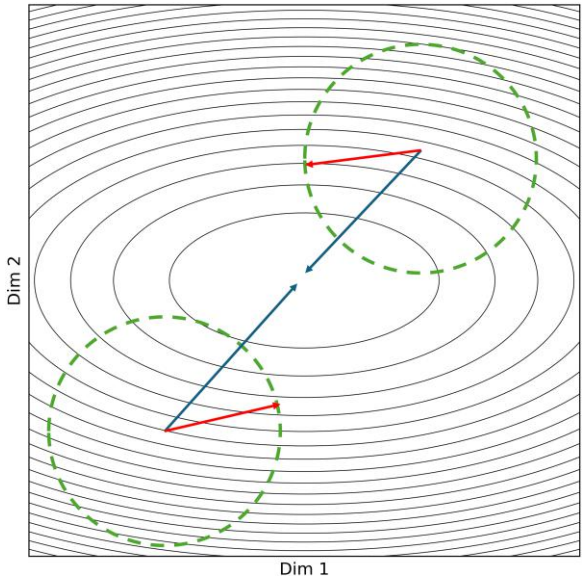
$$\overline{g}_t(x_i) = \frac{g_t(x_i)}{\max\left(1, \frac{\|g_t(x_i)\|_2}{C}\right)}. \tag{2}$$

2. **Noise injection.** After summing the clipped gradients within the minibatch, inject Gaussian noise (indicated by the red arrows in Figure 2). Instead of precisely following the blue arrow to the next parameter position, the update "wanders" in a randomly perturbed direction within the green circles in Figure 2. As Figure 2 suggests, these red arrows typically *misalign* with the true gradient direction; however, over many updates, the noise "cancels out," thereby preserving the overall gradient signal while obscuring the exact contribution of each individual. Formally, we average the clipped gradients and add Gaussian noise with variance $\sigma^2 C^2 \mathbf{I}$:

$$\tilde{g} = \frac{1}{|B|}\left(\sum_{i=1}^{N} \overline{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right). \tag{3}$$

Conceptually, this process is akin to adding a controlled amount of "measurement error" at each step. Each individual's representation might appear slightly "off" in the latent space, but because the misalignments are random—and the model trains across many individuals and many epochs—the aggregate direction retains the relationships among dimensions. Indeed, although DP-SGD can slow convergence or reduce accuracy somewhat compared with a nonprivate baseline, it preserves the *unbiasedness* of the learned model parameters.

Figure 2. A Two-Dimensional Illustration of How Differentially Private Stochastic Gradient Descent Operates on a Loss Surface



Notes. Each concentric curve represents an isocontour of the model's loss, whereas the green dashed circles denote the threshold for gradient clipping. After the raw gradient is clipped to this limit, Gaussian noise is injected (red arrows) before the final update step (blue arrows). By constraining and randomizing each gradient, DP-SGD restricts the impact of any single observation, thereby protecting individual privacy.

Parameters for norm clipping and added noise are chosen based on the desired privacy budget (ε, δ). A smaller ε indicates tighter privacy, necessitating larger noise or more aggressive clipping but also, potentially incurring greater accuracy trade-offs. Overall, by combining gradient clipping and stochastic noise, DP-SGD guarantees that altering or removing any single record x_i from the data set cannot significantly shift the learned parameters. Consequently, our latent space remains *differentially private* at every iteration, protecting sensitive data while still supporting a rich, generative framework.

Here, we have described DP-SGD in an oracle data set scenario. However, as we transition to more realistic settings—where must rely on common variables—the very same DP-SGD mechanism remains the bedrock of privacy protection. Thus, the next section extends the VAE to a multiencoder architecture that can *fuse* disparate data sets while retaining these privacy guarantees.

5.2. Realistic Scenario: Data Fusion to Crossimpute the Missing Data

In practice, an analyst does not have access to an oracle data set where all customers' survey and CRM records are fully observed. Instead, only a fraction of customers participate in the external survey, leading to gaps in the available data. A key challenge is to infer the missing CRM attributes for survey respondents and conversely,

to impute attitudes toward the service provider for a broader set of customers sampled from the CRM database. In this realistic setting, the survey and CRM data sets are completely disjoint, meaning that no individual appears in both. However, they share common variables, $\mathbf{x}_c$, which serve as a bridge between the two data sets. The goal is to leverage this shared information to recover the missing uncommon variables, $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, for each data set. Achieving this requires a shift in model architecture; a single encoder-decoder setup is no longer sufficient as it assumes full data availability. Instead, a multiencoder, multidecoder approach is necessary to process separate data partitions while still learning meaningful shared representations. Furthermore, real-world data exhibit complex structures that basic models may fail to capture. To improve the accuracy of crossimputation, we incorporate into the model normalizing flows, which not only allow for richer latent representations of intricate dependencies in the data partitions of each data set but also shall serve as the crosslearning (i.e., fusion) mechanism between data sets.

Although the internal validity of our architecture relies on the conditional independence assumption, external validity relies on the assumption that the data sets share sufficient distributional overlap in their common variables. Without such overlap in empirical settings, the model cannot reliably learn mappings between data sets for crossimputation. These two foundational assumptions are identical to prevalent data fusion approaches (Gilula et al. 2006), and advantageously, the introduction of differential privacy (Section 5.2.2) to our framework will necessitate no additional assumptions on the data-generating process.

To this end, we provide a sensitivity analysis in Appendix D on how overlap violations can affect reidentification risk in our empirical setting. We simulate these violations by removing subsets of customers from either the middle or the tail end of marginal distributions (w.r.t. specific attributes) and retrain PPDF under each scenario. The results show that reduced overlap between the surveys and CRM lowers reidentification risk overall, but subgroup effects vary sharply; removing tails can expose outliers and increase their reidentifiability, whereas removing midrange cases can blur distinctions and reduce reidentifiability. This highlights that the novelty of our framework lies not only in the model but also, in our proposed posterior reidentification metric (Section 5.3) that enables managers to assess customer-level reidentification risks and fine-tune their privacy budget accordingly.

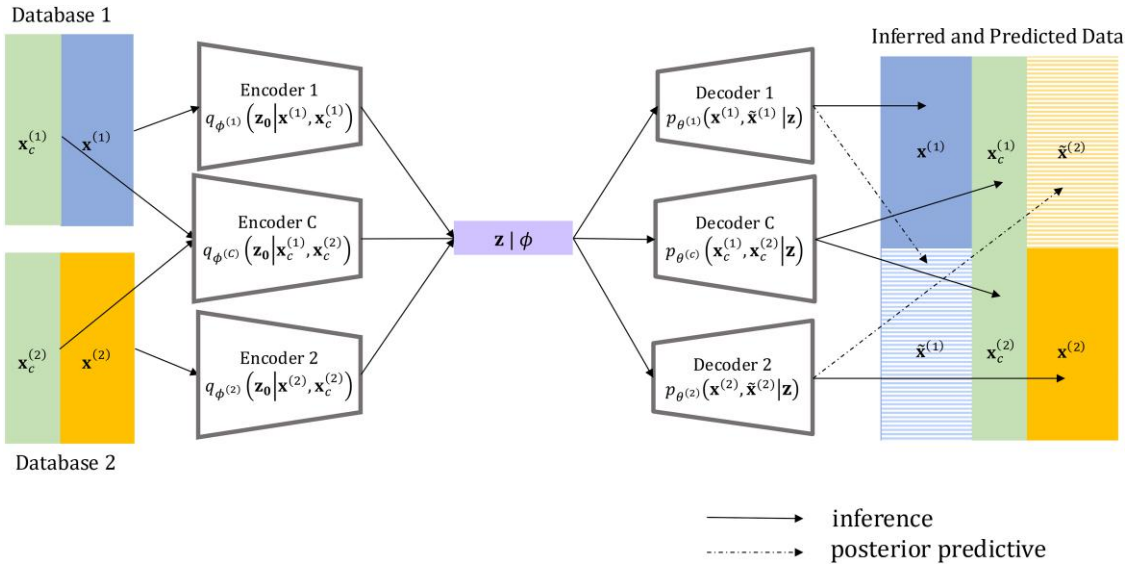
5.2.1. Multiencoder and Multidecoder Architecture for Data Fusion. Complementing its differentially private training, PPDF's multiencoder, multidecoder VAE architecture is motivated by the need for accurate learning of the component data sets in the fusion task, and it is a form of multimodal VAE. In the marketing literature, Tian et al. (2024) used a structured multimodal VAE architecture to learn joint embeddings from audiovisual and textual elements from TikTok content, and Sisodia et al. (2025) proposed a variant where a single-encoder network is paired with an image decoder as well as discriminative network on the corresponding structured data (i.e., marketing variables) to enable interpretable representation learning. In contrast to extant approaches, our approach handles multiple modalities (i.e., component data sets) that are not deterministically linked and must be probabilistically learned via common variables. This objective realizes data fusion's conditional independence assumption (Gilula et al. 2006) and manifests within our architecture as its own distinct set of encoder and decoder.

Figure 3 illustrates PPDF's architecture. At the core of our PPDF framework lies a set of three encoders, a single normalizing flow, and a set of decoders. Each encoder parameterizes variational densities for different data partitions; encoder 1 conditions on the common and uncommon variables of data set 1 ($\mathbf{x}_c^{(1)}$ and $\mathbf{x}^{(1)}$), encoder C conditions on the common variables across both data sets ($\mathbf{x}_c^{(1)}$ and $\mathbf{x}_c^{(2)}$), and encoder 2 conditions on the common and uncommon variables of data set 2 ($\mathbf{x}_c^{(2)}$ and $\mathbf{x}^{(2)}$). These encoders map the data to a common latent space ($\mathbf{z} = \{\mathbf{z}^{(1)}, \mathbf{z}^{(c)}, \mathbf{z}^{(2)}\}$). The model's three decoders enable the model to learn the joint distribution across data sets via posterior updating during the training phase. Posttraining, the model then undertakes a posterior predictive phase, indicated by dashed arrows in Figure 3, to impute via sampling from the posterior of the other data sets. This crossimputation serves to complete the "missing data" for each record in each data set conditional on common variables and thus, completing data fusion.

5.2.2. Privacy in the Data Fusion Context. The differentially private stochastic gradient descent mechanism remains unchanged from the illustrative single-data set scenario; it continues to ensure that no single observation disproportionately influences the learned model parameters. However, the shift to a data fusion setting introduces new challenges related to data set linkage and inference across disjoint partitions. In line with state-of-the-art work in this area, we implement a differentially private doubly stochastic variational inference algorithm for training in its (ε, δ) form (Prediger et al. 2022), which is derived from DP-SGD (Abadi et al. 2016).

5.2.3. Normalizing Flows to Improve Fusion Learning. One challenge of fitting VAEs is that they are limited in their ability to capture complex, real-world data-generating processes. Specifically, vanilla VAEs perform encoding using univariate Normal priors, $N(0, 1)$,

Figure 3. Detailed Architecture of the Data Fusion Process—A Trio of Encoders That Independently Parameterize Variational Densities for Different Subsets of Variables Within Each Data Set



Notes. At inference time, the VAE decodes the original data via three decoders into a self-reconstruction of the original data. The latent posterior $\mathbf{z}$ is then sampled and passed through the trained decoders 1, 2, and c to the crossimputation stage of the missing data. The solid arrows represents the *inference* step, whereas the dashed arrows represents the *posterior predictive* step. Differential privacy will be integrated through DP-SGD as will be illustrated in Figure 4.

on the latent variables, in part motivated by the tractability in forming the loss function.

We extend the expressiveness of the latent encoding with normalizing flows, serving not only to enhance model flexibility but also, to manage the complexity that comes to the forefront during data fusion. Moreover, the incorporation of normalizing flows serves as the mechanism of learning a joint distribution across data sets; despite being assumed independent *in the prior*, the latent variables $\mathbf{z}$ across the data sets share a common distribution *in the posterior*. That is, each data set maps to its own latent representation—denoted here as $\mathbf{z}_1$, $\mathbf{z}_2$, and $\mathbf{z}_c$—but these are transformed through shared flows into a common latent space, allowing samples from one data set's latent code to be used for generating or imputing data in another (see Online Appendix EC.7). Specifically, this posterior alignment in $\mathbf{z}$ implies that we can validly sample the following:

- $p_{\theta^{(1)}}(\mathbf{x}^{(1)}|\mathbf{z}_1)$ and $p_{\theta^{(c)}}(\mathbf{x}_c^{(1)}|\mathbf{z}_c)$ (i.e., the self-reconstruction of data set 1),
- $p_{\theta^{(2)}}(\mathbf{x}^{(2)}|\mathbf{z}_2)$ and $p_{\theta^{(c)}}(\mathbf{x}_c^{(2)}|\mathbf{z}_2)$ (i.e., the self-reconstruction of data set 2),
- $p_{\theta}(\tilde{\mathbf{x}}^{(1)}|\mathbf{z}_2)$ (i.e., the crossimputation of data set 1 (w.r.t. data set 2)), and
- $p_{\theta}(\tilde{\mathbf{x}}^{(2)}|\mathbf{z}_1)$ (i.e., the crossimputation of data set 2 (w.r.t. data set 1).

Taken together, this results in the "fused" joint density $p(\mathbf{x}^{(1)}, \tilde{\mathbf{x}}^{(1)}, \mathbf{x}^{(2)}, \tilde{\mathbf{x}}^{(2)}, \mathbf{x}_c^{(1)}, \mathbf{x}_c^{(2)})$. It is worth noting that the above marginal posteriors expressed in terms of a two-data set data fusion can be arbitrarily extended to J-data sets without loss of generality to its privacy budget (see Lemma 2) and limited only by computational resources. Ultimately, data fusion's goal is that we can condition on any one data set's common variable and generate samples of itself as well as impute those of the other. As such, we remind readers of the theoretical results of Lemma 2 (Section 4.2); the over-and-above privacy risks of data fusion arise from the additional crossimputation "models" (i.e., posterior marginals), which do not exist in generative models of individual data sets.

5.3. Posterior Reidentification Metric

Echoing the vignette from Section 4.1, we define the reidentification risk of individual i in data fusion as the closeness between their true values in one data set and the values predicted for them using the other data set, where predictions are made via crossimputation conditional on the common variables. The idea is that the combination of sufficiently accurate crossimputations can triangulate onto a user's record over and above others, and should the target data set contain PII (e.g., CRM), then that user's identity is *reidentified*.

Building on this definition, we propose a user-level reidentification risk metric tailored to data fusion. This metric is designed to go beyond the abstract guarantees offered by differential privacy parameters (ε, δ) and instead, satisfy three practical criteria. (1) It leverages the uncertainty estimates from PPDF's posterior to support actionable decisions by managers, such as identifying which users are at highest risk of reidentification or

making informed trade-offs between predictive accuracy and reidentification risk. (2) It quantifies reidentification risk in finite samples while allowing it to converge to zero as the data fusion sample size increases. (3) It remains usable even when firms lack access to ground-truth validation data, such as internal surveys that cannot be externally verified.

To operationalize this user-level reidentification metric in practice, we distinguish between two scenarios based on data availability. First, in settings where a firm possesses a ground-truth validation data set—such as an internal survey that overlaps with CRM data—we can directly compare predicted and actual values to assess reidentification risk. Second, in more common situations where such validation data are unavailable, we introduce a tolerance-based proxy that estimates risk based on how uniquely identifiable a user's attributes are relative to others in the data set. We detail each approach below.

5.3.1. Ground-Truth Scenario. Consider the telecom context where PII is found in the CRM and the carrier possesses the ground-truth validation survey. Let $\tilde{\mathbf{x}}_i^{(\text{CRM})}$ denote the posterior predictive crossimputations of user i's CRM variables conditional on his or her survey (common) variables:

$$\tilde{\mathbf{x}}_i^{(\text{CRM})} \sim p(\mathbf{x}_i^{(\text{CRM})} | \mathbf{x}_i^{(\text{Survey})}). \tag{4}$$

We define survey respondent i as being reidentified in the CRM data if their crossimputed CRM attributes ($\tilde{\mathbf{x}}_i^{(\text{CRM})}$) have the lowest error with respect to their ground-truth CRM attributes ($\mathbf{x}_i$) when compared against the CRM attributes $\mathbf{x}_j$ of all $j = 1, \ldots, N$ CRM customers, which we operationalize in terms of cosine similarity $\mathcal{D}_{i,j} \in [0, 1]$:

$$\mathcal{D}_{i,j} = \frac{\mathbf{x}_j^{(\text{CRM})\top} \tilde{\mathbf{x}}_i^{(\text{CRM})}}{\|\mathbf{x}_j^{(\text{CRM})}\| \, \|\tilde{\mathbf{x}}_i^{(\text{CRM})}\|}. \tag{5}$$

To quantify the privacy risk of an individual i, we then calculate $j = \text{argmin}_j \mathcal{D}_{i,j}$. The central idea is that if $i = j$, then respondent i has been reidentified. Conversely, should $i \neq j$, then the more individuals j who have lower $\mathcal{D}_{i,j} < \mathcal{D}_{i,i}$, the more indistinguishable i is, and thereby, the lower the reidentification risk is for i.

Hence, innate to this metric is the inductive bias that as N goes to ∞, reidentification risk is expected to go to zero for all users, even without privacy-enhancing methods. However, in reality, most real-world survey and CRM data sets—including those used by the telecom carrier in Section 6—exist in finite samples and do not benefit from this asymptotic guarantee. For these typical data regimes, we will show that it is more *economical* to utilize PPDF to induce privacy than to collect more data. In doing so, we will disentangle the incremental efficacy of PPDF for data sets of different sizes and showcase the intuitiveness of the proposed reidentification metric.

To capture the inherent uncertainty in data fusion and its impact on reidentification, let $p(\tilde{\mathbf{x}}_i | \mathbf{x}_i^{(*)})$ denote the posterior predictive distribution of individual i's crossimputations from any $(*)$ counterpart data set (e.g., CRM variables crossimputed from survey variables). The full posterior distribution of individual i's reidentification metric is then given by

$$p(\mathcal{D}_{i,j} | \mathbf{x}_j, \mathbf{x}_i^{(*)}) = p(\mathcal{D}_{i,j} | \mathbf{x}_j, \tilde{\mathbf{x}}_i) \cdot p(\tilde{\mathbf{x}}_i | \mathbf{x}_i^{(*)}). \tag{6}$$

To quantify the privacy risk of individual i using Equation (6), we use quantile differences in the distribution of distances $p(\mathcal{D}_{i,i} | \mathbf{x}_i, \mathbf{x}_i^{(*)})$ to $p(\mathcal{D}_{i,j} | \mathbf{x}_j, \mathbf{x}_i^{(*)})$, $\forall j \neq i$. We define the "privacy index" $N_i(\alpha)$, which describes the number of individuals for whom the distance to individual i's reconstructed data is less than or equal to a target quantile α (e.g., 1st, 5th, 95th, and 99th quantiles),

$$N_i(\alpha) = |\{j : q_{\alpha, p(\mathcal{D}_{i,j})} < q_{\alpha, p(\mathcal{D}_{i,i})}\}|, \tag{7}$$

where, $q_{\alpha, p(\mathcal{D}_{i,j})}$ denotes the α-quantile of $\mathcal{D}_{i,j}$'s posterior distribution. The quantity $N_i(\alpha)$ measures how many other individuals j have an α-quantile distance that is *smaller* than i's own α-quantile distance $q_{\alpha, p(\mathcal{D}_{i,i})}$ (i.e., calculated from posterior predictive draws). In other words, a higher value of $N_i(\alpha)$ means that more individuals are at least as good a match to i's crossimputed data as i is to its own data—making i relatively less unique and less likely to be reidentified.

5.3.2. No Validation Data Set Scenario. Recognizing that not all firms readily have an analogue of the internal validation survey at their disposal to implement the above framework, we now propose an alternative operationalization based on *tolerance* over $\mathcal{D}_{i,j}$.

Per Equation (6), let $T_i(\tau)$ indicate the number of individuals in the target data set whose common-variable crossimputations ($\tilde{\mathbf{x}}_j$) fall within a prespecified tolerance τ of individual i's observed ($\mathbf{x}_i$) in the counterpart data set,

$$T_i(\tau) = |j : \mathcal{D}_{i,j} \geq 1 - \tau|, \tag{8}$$

where $\tau \in [0, 1]$ is a threshold indicating the acceptable similarity margin and is the nonground-truth analogue to the quantile designation used in Equation (7). Hence, rather than requiring an explicit ground-truth record, $T_i(\tau)$ quantifies how many other users exhibit a desired level of similarity to individual i after data fusion, especially in the data set where PII exist. Larger values of $T_i(\tau)$ imply lower reidentification risk as more individuals are similar to individual i. This tolerance-based approach provides firms a practical metric for managing reidentification risk even in the absence of ground-truth validation data, facilitating intuitive managerial

decisions about privacy-accuracy trade-offs inherent in data fusion scenarios.

Taken together and represented visually in Figure 4, PPDF is an end-to-end privacy-preserving framework for data fusion—from model to benchmarking—across the prototypical CRM analytical workflow, of which we will empirically demonstrate in the next section.

6. Application: Privacy-Preserving Targeting and Heterogeneity Analysis

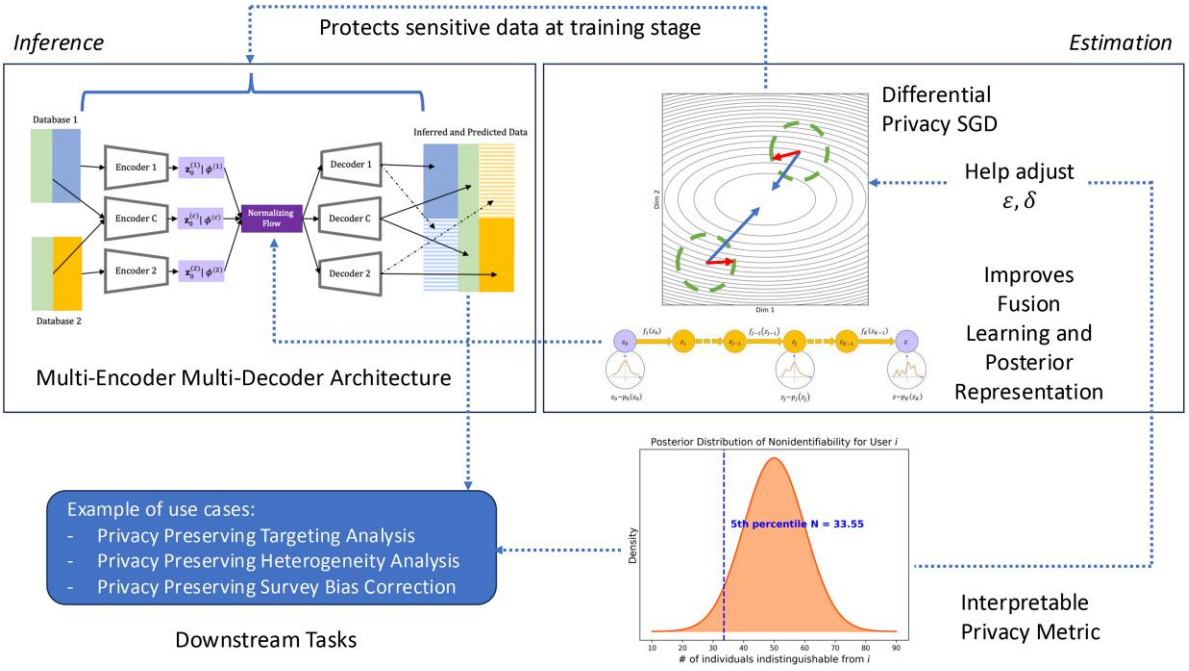
We now illustrate the utility of PPDF using the telecom data. After establishing the analytical bounds on data fusion's unique privacy risks (Section 4) and outlining the proposed methodology (Section 5), this section seeks to empirically address two questions delineated by stakeholder groups. First, for managers in practical CRM scenarios, does there exist a feasible privacy-accuracy trade-off when utilizing PPDF? And if so, who among the customers are most at risk for reidentification and thereby, nontrivially benefited from the privacy gains afforded by PPDF (as opposed to privacy gained organically because of data size, etc.)? On the first question, we illustrate a prototypical churn-prevention campaign that leverages data fusion to improve predictive accuracy, with the goal of establishing a *feasible* privacy budget—defined as one that begets zero reidentifications while attaining competitive predictive accuracy. The idea is that managers are much more likely to proactively adopt a privacy-enhancing technology, such as PPDF, when the efficacy of their workflow is unchanged. Then, honing in on the feasible privacy budget found for the churn campaign, we perform an accompanying heterogeneity analysis to illuminate the composition of the most at-risk customers by contrasting against a version of PPDF tuned to no privacy guarantees (i.e., budget). We now discuss the campaign setting and targeting rules at the telecom carrier.

6.1. Campaign Setup

We consider a hypothetical targeting campaign aimed at preventing customer churn in a cost-effective manner. In this setting, we observe two critical outcome variables for each customer: (i) *trailing 12-month revenue (TTM)*, capturing the individual's recent spending level, and (ii) a binary indicator for whether the customer *will have churned* in the next 12 months. Our primary interest is in designing a targeting strategy based on these outcomes—focusing on customers at greater risk of churning who also exhibit meaningful revenue potential.

Although the specific methods for evaluating campaign performance are discussed in the next subsection, we highlight here that the core problem involves

Figure 4. Privacy-Preserving Data Fusion Framework



Notes. This figure illustrates the workflow of the PPDF framework, which enables secure data fusion while maintaining privacy guarantees. The multiencoder, multidecoder architecture performs the fusion of disparate data sets through a normalizing flow-based generative approach. Differentially private stochastic gradient descent (SGD) ensures privacy during training by adjusting the privacy parameters ε, δ. The framework introduces an interpretable privacy metric, quantifying the risk of reidentification through posterior distributions of nonidentifiability. Example use cases include privacy-preserving targeting analysis, heterogeneity analysis, and survey bias correction, demonstrating the broad applicability of PPDF in privacy-sensitive data fusion scenarios.

balancing the benefits of retaining high-value customers against the costs of targeting those who will churn in any case. This scenario provides a clear environment for exploring how data fusion can improve targeting decisions as it involves enriching churn predictions with additional survey information not originally available in the CRM. However, as we shall see, such enrichment will also coincide with undesirable reidentification of respondents from the anonymous surveys. Hence, in addition to the primary targeting objective, we analyze how customers' survey responses and CRM engagement patterns jointly shape their likelihood of churning and their risk of reidentification. Specifically, we aim to uncover whether certain subpopulations—defined by distinct combinations of attitudinal and behavioral attributes—are systematically more sensitive or more prone to reidentification.

6.2. Alternative Approaches

6.2.1. XGBoost. In this targeting exercise, one fundamental question is whether we can accurately predict which customers are more likely to churn in the next 12 months. A standard approach in many firms is to use a predictive model using the data readily available in the CRM system. This model then serves to rank customers according to churn risk and guide targeting decisions.

We use XGBoost (Chen and Guestrin 2016) as the candidate downstream discriminative model for three reasons. The first is its predictive performance because gradient-boosted trees tends to produce state-of-the-art accuracy for a wide range of tabular data tasks. The second is scalability; XGBoost is computationally efficient, allowing it to handle large data sets and complex sets of variables. The third is its widespread use in practice. Additionally, XGBoost is fitted directly on $\mathbf{x}_i^{(\mathrm{CRM})}$ to serve as a "sanity check" benchmark.

6.2.2. Multiencoder and Multidecoder Without DP-SGD. One way to fuse the CRM and survey data sets—and thus, leverage both sets of variables in predicting churn—is via a multiencoder, multidecoder variational autoencoder as described in Section 5.2.1. However, unlike our proposed PPDF approach, where training occurs via DP-SGD, this comparison variant omits the privatized training. The advantage is that the model does not have to inject noise into its gradient computations, potentially leading to more accurate data reconstruction. Yet, it lacks any formal privacy guarantee, exposing the fused data to higher risks of reidentification.

6.2.3. Local PPDF. PPDF makes use of *global* differential privacy, in which the privacy budget is realized and implemented during the fusion process (see Section 5.1.2). An alternative strategy is *local* differential privacy. Local DP (Kasiviswanathan et al. 2011) involves running differentially private learning of the data sets separately and independently such that any subsequent data fusion steps automatically inherit the privacy guarantees by DP's composability property (see Section 4.2). The downside is the amount of noise generated in its piecewise workflow, which can offset the downstream model accuracy. Nevertheless, local DP serves as an important benchmark for PPDF as in the spirit of *feasibility*, it could very well be "good enough." Local DP would simplify the steps for the stakeholders; the data owner perturbs the data and sends them to the data curator, who then uses a nondifferentially private version of PPDF (i.e., performs *local PPDF*).

In the context of data fusion, local DP adds noise to each of the data sets prefusion by outputting the true value for a given consumer with a certain probability that depends on privacy budget ε and outputs a randomized value otherwise. In the telecom context, as all variables are categorical, we utilize the generalized randomized response (GRR) technique (adequate for discrete data) (Kairouz et al. 2016). Given a user i with value k in the set of D possible true values, denoting $\hat{r}$ the random variable that represents the response of a user i, the GRR outputs

$$P(\hat{r}=k)=\begin{cases}\frac{\exp(\varepsilon)}{\exp(\varepsilon)+D-1} & \text{if } r=k\\ \frac{1}{\exp(\varepsilon)+D-1} & \text{if } r\neq k\end{cases}. \tag{9}$$

In comparing our proposed "global" PPDF with local PPDF and non-DP baselines, we find that "global" PPDF achieves consistently higher privacy protection (zero unique reidentifications) and maintains strong data reconstruction accuracy across all tested privacy budgets. By contrast, local PPDF exhibits a more variable trade-off between accuracy and privacy, and the non-DP benchmark—although most accurate—suffers from significant reidentification risks. Full details of this comparison, including goodness of fit and robustness checks via simulation, are provided in Appendix A.

6.3. Metrics for Performance

This section describes how we assess each model's ability to (i) accurately place individuals into the correct cells of a three-way frequency table involving *churn*, *TTM*, and *LTR* and (ii) protect privacy via differential privacy parameters. We then operationalize the accuracy versus privacy trade-off in terms of campaign lift and posterior reidentification risk.

6.3.1. Churn, TTM, and LTR. We define *trailing 12-month revenue* at the individual level as a normalized percentile measure of spending over the past year. *Churn* is a binary indicator of whether a customer cancels their service in the next 12 months. We combine these variables into a hypothetical campaign objective function by comparing the retained revenue against the cost of targeting.[8] Let π_i be the net profit outcome, which is defined

as an individual i's TTM minus the (fixed) cost of targeting them. The overall *campaign lift* is then

$$\pi_{\text{LIFT}} = \frac{\sum_{i=1}^{N} \pi_i}{\sum_{i=1}^{N} \text{TTM}_i}.$$

We further consider conditional expected lifts, such as $\mathbb{E}[\pi_{\text{LIFT}}|\text{Churnonly}]$, $\mathbb{E}[\pi_{\text{LIFT}}|\text{Churn, TTM}]$, $\mathbb{E}[\pi_{\text{LIFT}}|\text{Churn, LTR}]$, and $\mathbb{E}[\pi_{\text{LIFT}}|\text{Churn, TTM, LTR}]$, where tighter targeting criteria can be imposed if, for example, the firm wishes to reach only high-spend churners or those with certain LTR values.

6.3.2. Three-Way Frequency Table. Each customer is classified by whether they churn, their TTM quantile, and their LTR category. Placing a customer in the "correct" cell of this table (e.g., predicting both their churn status and TTM ranking accurately) yields higher lift by minimizing false positives (nonchurners who get targeted) and false negatives (churners who are overlooked). Model accuracy effectively determines how well the campaign can allocate resources to profitable segments.

6.3.3. Privacy Measures. We quantify privacy via the privacy budget ε and the posterior reidentification metric. Reducing ε tightens privacy but generally increases noise during training, potentially lowering predictive accuracy. The posterior reidentification metric indicates how many individuals can be uniquely matched across fused data sources; smaller ε typically reduces unique matches, yet too small of a privacy budget, meaning less reidentification risk, may degrade the campaign's overall lift.

6.3.4. Trade-off: Campaign Lift vs. Reidentification Risk. A typical managerial concern is to evaluate the net gain from a more accurate predictive model at the cost of higher reidentification risk. If, for instance, decreasing the privacy budget ε from 2.0 to 0.5 successfully cuts the reidentification rate in half but reduces the expected campaign lift by only a few percentage points, then stricter privacy might be appealing. Conversely, if lowering ε imposes a large drop in net lift, the firm may accept a slightly higher reidentification risk in exchange for robust campaign performance. Through this lens, each model's predictions are evaluated not only on $\mathbb{E}[\pi_{\text{LIFT}}]$ but also, on how changes in ε translate into more or fewer uniquely reidentified individuals.

6.4. Results

6.4.1. Predicted Targeting. The discussion now turns to using PPDF for the focal targeting exercise, demonstrating that it is possible to achieve competitive campaign lift while ensuring user privacy. The model-based targeting is operationalized specifically as a downstream exercise. Conditioning on the uncommon and common CRM variables $X_i^{(\text{CRM})}$ of the customers, the steps are as follows:

1. drawing posterior predictive samples of the joint "CRM-and-survey" distribution $p\left(\tilde{\mathbf{x}}_i^{(\text{CRM})}, \tilde{\mathbf{x}}_i^{(\text{Survey})} | \mathbf{x}_i^{(\text{CRM})}\right)$ trained on the CRM database and the external, anonymous survey;
2. using these samples as input variables, training downstream discriminative models for churn, TTM, and LTR as outcomes; and
3. using predictions on customers' churn, TTM, and LTR to run the targeting campaign on the holdout/test sample (composed of the ground-truth internal validation survey), employing the same cutoff-based targeting logic previously discussed.

Detailed model comparisons are provided in Appendix C, including alternative data fusion methods. The overall best discriminative performance is observed in PPDF trained *without* differential privacy (i.e., no privacy budget), which notably overlaps or even exceeds XGBoost trained directly on $\mathbf{x}_i^{(\text{CRM})}$ across the 99% highest posterior density (HPD). As shown in Table 2, broadly, maximal campaign lift across all models is observed when segmenting customers in terms of both their predicted TTM and LTR (see the "Churn, TTM, and LTR" column), which indicates that data fusion of the survey to CRM indeed enhances downstream performance. This is in line with the work of Ascarza (2018), which highlighted the importance of targeting based not solely on the likelihood to churn but rather, also on the likelihood of responding to targeting campaigns. Having said that, this outcome can be seen as a contradiction to Rafieian and Yoganarasimhan (2021), who show that ad revenue is maximized when the ad network does not allow advertisers to engage in behavioral targeting. With PPDF, the lift derived from using both behavioral and attitudinal data is large and significant. Leveraging the ground-truth validation survey data set, our results on the value add of data fusion for model-based targeting are reaffirmed via model-free evidence, details of which can be found in Online Appendix EC.1.2.

In considering the trade-off between accuracy and privacy, we find that variations in PPDF's privacy budget result in statistically and economically significant differences in estimated campaign lift. Notably, the oft-cited (Prediger et al. 2022) privacy budget of $\varepsilon = 1.0$ is managerially suboptimal (i.e., maximum of 0.84% lift). Yet, even at the lowest tested privacy budget ($\varepsilon = 5.0$), it remains a fact that no survey respondents are reidentified (calibrated to the fifth percentile of the individual-level posterior reidentification metric) while simultaneously (1) achieving campaign lift indistinguishable from that of XGBoost[9] and (2) only 0.2 percentage points shy of the estimated lift from a benchmark variant of PPDF trained

Table 2. Estimated Lift of Model-Based Churn Targeting: PPDF vs. Nonprivatized and Privatized Benchmarks

Model	Privacy budget	Churn only, %	Churn and TTM, %	Churn and LTR, %	Churn, TTM, and LTR, %
PPDF	None	0.77 [0.76–0.78]	1.36 [1.33–1.39]	1.26 [1.24–1.29]	1.66 [1.63–1.68]
XGBoost	None	0.64 N/A	1.19 N/A	1.20 N/A	1.45 N/A
PPDF	$\varepsilon = 5.0$	**0.68** **[0.66–0.70]**	**1.20** **[1.15–1.25]**	**1.16** **[1.11–1.20]**	**1.46** **[1.41–1.51]**
PPDF	$\varepsilon = 2.0$	0.59 [0.57–0.61]	0.58 [0.52–0.64]	0.46 [0.39–0.52]	0.84 [0.77–0.90]
PPDF	$\varepsilon = 1.0$	0.53 [0.50–0.55]	0.49 [0.43–0.54]	0.61 [0.55–0.66]	0.84 [0.78–0.90]
PPDF	$\varepsilon = 0.50$	0.45 [0.42–0.49]	0.44 [0.37–0.50]	0.51 [0.44–0.57]	0.72 [0.65–0.79]
PPDF	$\varepsilon = 0.25$	0.47 [0.45–0.50]	0.45 [0.39–0.51]	0.34 [0.28–0.41]	0.65 [0.59–0.72]

Notes. The 99% highest posterior densities are in brackets. $\varepsilon = 0.25$ is most stringent budget, and $\varepsilon = 5.0$ the least stringent budget. PPDF trained under $\varepsilon = 5.0$ achieves comparable campaign lift to XGBoost (i.e., nonprivacy-preserving discriminative model trained directly on CRM data). All shown PPDF privacy budget variants result in zero reidentifications (i.e., no individuals were reidentified when using PPDF), whereas the PPDF variant without privacy results in 6.98% of validation survey respondents reidentified (see Figure 5). N/A, not applicable. Bold indicates best performing model variable.

without DP (i.e., no privacy budget and best overall lift). Examples of user-level reidentification metric plots across privacy budgets can be found in Online Appendix EC.3.

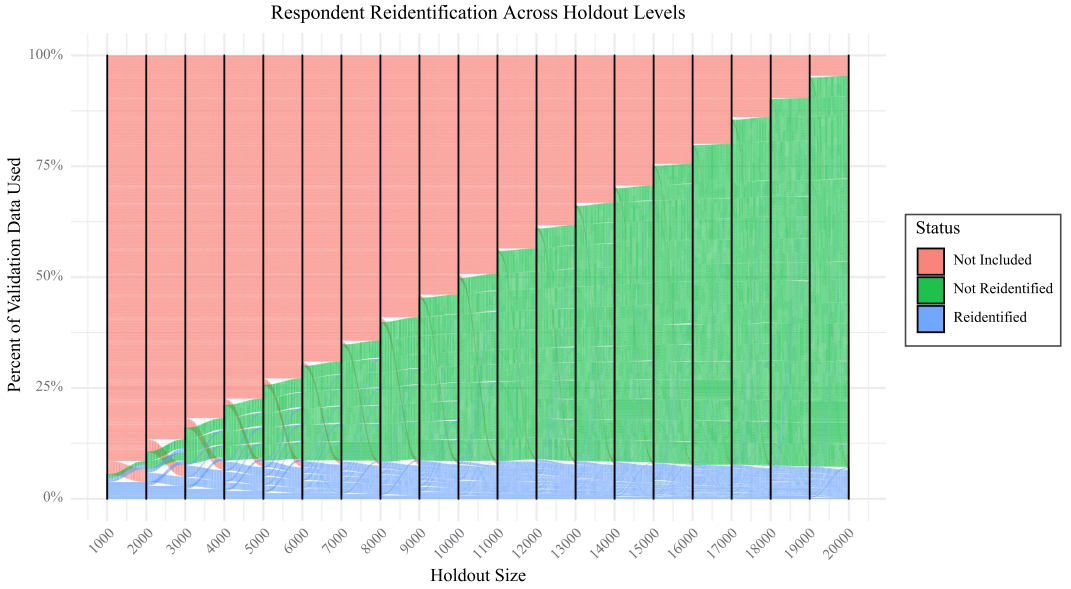
As noted earlier, managers are much more likely to adopt PPDF when their task efficacy is unchanged, and simultaneously, they can be ensured of zero reidentifications under data fusion. Moreover, should the firm be willing to adopt a less stringent privacy policy (i.e., zero reidentifications at the 10th percentile or even the median of the individual-level posterior reidentification metrics), then PPDF would be expected to exceed XGBoost in performance and converge onto the lift found under PPDF without DP training (max lift of 1.66%). Our objective, rather, is to demonstrate the *feasibility* of attaining zero reidentifications (privacy) and equivalent performance to existing practices (accuracy).

Taken together, although our results are context specific, they empirically corroborate the notion that PPDF's privacy guarantees extend to downstream tasks as a result of differential privacy's sequential composability property (Dwork et al. 2006a, b).

6.4.2. Heterogeneity Analysis. In this subsection, we explore the core questions regarding individuals who become reidentified under the nonprivate version of our fusion model and how they fare when using the differentially private version of our model (tuned to $\varepsilon = 5.0$). First, we examine the number of reidentified survey respondents across different holdout sample sizes utilizing the ground-truth internal survey, from which we draw varying holdout sample sizes from 1,000 to 20,000 individuals (the remainder is simply excluded from this analysis). This serves as the function of assessing the organic increase in privacy because of data set size as opposed to PPDF by increasing the a priori chance of having others who may be similar to a target customer. Yet, we find a sizable group of customers who are persistently reidentified, despite the increase in holdout size. Figure 5 represents an alluvial diagram of reidentification status as the holdout sample size increases. The maximal number of individuals (1,773) reidentified was achieved at a sample size of 9,000. Yet, at 20,000, 1,396 individuals (6.98%) remain reidentified—of whom 90.1% are identical to those at the maximum. As noted in the discussion on the reidentification metric (Section 5.3), we believe that as the data set size approaches infinity, no respondent would be reidentified organically. However, the results suggest a finite sample reality for managers; PPDF readily ensures zero reidentifications utilizing the data at hand, whereas collecting more data to attain the same level of privacy preservation would likely be more time consuming, logistically complex, and less economical.

Second, we show that drivers of individual-level reidentification when using the nondifferentially private model can be predictable and in turn, that privacy risk is heterogeneous. We run post hoc binary logit models, where the dependent variable flags the individuals who are reidentified at a holdout sample size of 20,000. The independent variables are the common variables, the uncommon CRM variables and the uncommon survey variables. The full results are in Online Appendix EC.8. Most of these variables are significant predictors of reidentification (whereas signs are less meaningful as a negative coefficient simply means

Figure 5. Alluvial Diagram Illustrating the Reidentification Status of Validation Survey Respondents from Data Fusion-Based Targeting Exercises (PPDF Trained *Without* Differential Privacy per Table 2) Across Holdout Sample Sizes (from 1,000 to 20,000 Individuals)



Notes. Respondents are categorized as not included at the holdout level (red), included but not reidentified (green), or reidentified (blue). The maximal number of individuals reidentified (1,773) occurs at a sample size of 9,000. At largest holdout size (20,000), 1,396 individuals (6.98%) remain reidentified, with 90.1% overlapping with those reidentified at the maximum.

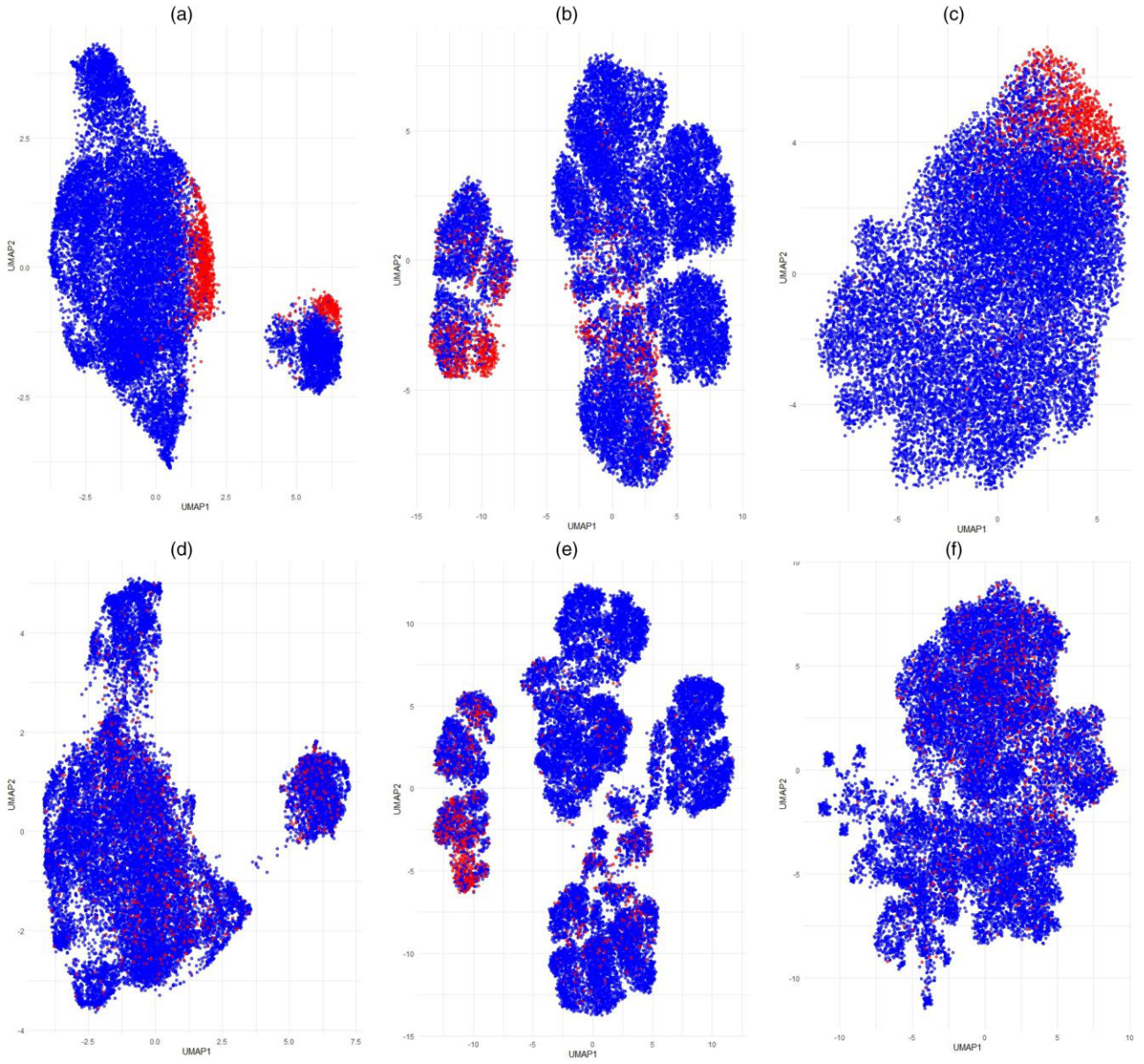
that lower values increase reidentification). As survey responses become extreme or out of range in a way that makes the respondents unique, the probability of being reidentified to their PII in the CRM goes up under non-DP data fusion. However, what this analysis does not tell us is whether those who are reidentified represent clusters onto themselves or whether they are edge cases of broader customer clusters.

Hence, third, we investigate any clustering patterns among the reidentified individuals. To this end, we undertake an unsupervised clustering analysis on all 20,000 holdout individuals—using the common variables, the uncommon CRM variables, and the uncommon survey variables—but excluding reidentification labels. Then, post hoc, we superimpose the reidentification labels to investigate whether discernible patterns emerge. Because of our large sample size and the large number of categorical variables, we apply dimensionality reduction using UMAP.[10] Two-dimensional visualization of the data is shown in Figure 6; panels (a)–(c) of Figure 6 visualize the normalized data when the nondifferentially private version of our model is used, and panels (d)–(f) of Figure 6 visualize the normalized data when the differentially private version is used. We break down the data by types of variables (CRM uncommon, survey uncommon, and common variables). Post hoc, we color in red in Figure 6 the individuals who are reidentified among the holdouts, and we color in blue in Figure 6 those who are not. As before, the privacy budget $\varepsilon = 5.0$ is used as calibrated from the churn-targeting exercise. Crucially, note that in the full model (i.e., DP-SGD training) (panels (d)–(f) of Figure 6), respondents who otherwise would be reidentified (i.e., red dots in Figure 6) are now interspersed throughout their respective cluster, whereas they formed clear "edge case" subgroups when differential privacy is removed (i.e., panels (a)–(c) of Figure 6). Moreover, note that despite such interspersion, the overall configuration of clusters within the Uniform Manifold Approximation and Projection (UMAP) domain is sustained across panels (a)–(c) of Figure 6 (no DP) to panels (d)–(f) of Figure 6 (DP). This finding epitomizes the earlier schematic illustration (Figure 2) of the mechanisms and advantages of PPDF's architecture and training; it induces privacy by obfuscating the learned embeddings of individuals but doing so while preserving the higher-level geometries in the embedding space. It is this explicit and simultaneous balancing of these two opposing needs in which PPDF was able to attain a privacy budget that also exhibited competitive accuracy—suggesting that managers can feasibly apply PPDF to their existing customer-level data fusion workflows.

7. Discussion

Reidentification risk arises whenever models are trained on fused data sets that combine information about individuals across sources—even if no single data set is ever

Figure 6. Clustering Patterns Among Holdout Individuals Without (Panels (a)–(c)) and Under (Panels (d)–(f)) Differential Privacy ($\varepsilon = 5.0$)



Notes. Red points indicate individuals who were reidentified in the no DP scenario. Under DP, none of these same individuals are reidentified, but they are colored red for comparison. DP "pushes" prior outliers into denser areas, dispersing them enough to thwart reidentification. (a) No DP: CRM uncommon. (b) No DP: common variables. (c) No DP: survey uncommon. (d) DP ($\varepsilon = 5.0$): CRM uncommon. (e) DP ($\varepsilon = 5.0$): common variables. (f) DP ($\varepsilon = 5.0$): survey uncommon.

directly exposed and even if there is no intent to target or identify anyone. The mere act of learning statistical relationships between data sets can enable unintended inferences; given partial information about a person (e.g., survey responses), a trained model may reveal hidden attributes from another source (e.g., CRM records). This risk persists even when one of the data sets remains offline or unused in deployment. To eliminate this model-based inference vector, our approach—PPDF—incorporates differential privacy at training time, ensuring that no individual record leaves a detectable trace in the model while preserving the structure necessary for effective prediction and decision making.

In evaluating PPDF, our primary focus lies in its ability to mitigate reidentification risks while delivering strong inferential performance. Yet, at its core as a probabilistic data fusion model (Gilula et al. 2006), PPDF's architecture also effectively handles missing data and

selection bias using common variables to establish conditional independence, accommodating both missing at random and missing completely at random mechanisms. As detailed in Online Appendix EC.1, we show that the telecom carrier's survey respondents exhibit systematic differences from the broader customer base; specifically, customers with more extreme attitudes (providing the smallest and highest values on scale points) were disproportionately likely to respond to the survey. This affirms commonly observed patterns in online reviews and customer feedback systems (Schoenmueller et al. 2020). Given its superior performance against benchmark data fusion methods (Table C.1 in Appendix C), PPDF represents an expressive and scalable data fusion method independent of its privatized training.

Turning to our predicted targeting exercise, the highest campaign lift (1.66%) emerges under the nonprivate variant of PPDF, which still falls short of the 7.84% derived from a purely model-free baseline (see Online Appendix EC.1.2). This gap primarily stems from a simplistic cutoff-based targeting rule, underscoring how more refined discriminative approaches (e.g., Verhoef 2003, De Haan et al. 2015, Lemmens and Gupta 2020) could further enhance lift when integrated with PPDF. Rather than chasing top-tier profitability estimates, our goal is to demonstrate that privacy-respecting data fusion can achieve competitive performance in routine managerial tasks. In the telecom context, there appears little reason not to adopt differential privacy within this framework.

Our findings also highlight how conventional guidelines from the literature and practice on setting $\varepsilon < 1.0$ can be ill suited to marketing applications. Indeed, larger budgets (e.g., $\varepsilon \approx 5.0$) deliver reasonable privacy protection—verified by our posterior reidentification metric—and maintain revenue-generating accuracy in targeting. However, pinpointing an "optimal" privacy budget requires richer field data to quantify the marginal dollar value of privacy. Moreover, we used a backward-looking TTM proxy and presumed guaranteed retention upon targeting, simplifying any interaction effect between privacy and campaign effectiveness (Ascarza 2018).

In addition, smaller or more vulnerable subpopulations (such as rural or minority groups) could experience disproportionate data obfuscation (i.e., accuracy loss) under stricter privacy regimes (Ruggles 2024). In turn, as data fusion rests upon assuming sufficient data overlap (see Section 5.2), violations of distributional overlap can have complex—and even contradictory—ramifications on the reidentification risks of these subpopulations as we show in Appendix D. Future investigations could explore how to safeguard distinct groups without asymmetrically eroding their predictive accuracy. Because our methodology adapts to a broad range of marketing tasks, PPDF's scope extends beyond churn prevention. For instance, multiple firms could collaborate to learn market share patterns or product complementarities from partitioned databases without fully relinquishing private information. PPDF likewise enables a firm to split sensitive data and store them in separate silos, limiting the harm from data breaches by ensuring that PII is never colocated with attitudinal or behavioral histories. These applications showcase the potential for robust analytics and privacy to coexist, thereby fostering greater trust in data-driven marketing practices.

8. Summary

This research introduces a privacy-preserving data fusion framework that enables practitioners to fuse multiple data sets securely while retaining the analytical power of data fusion techniques. By combining variational autoencoders, normalizing flows, and differential privacy, PPDF significantly reduces reidentification risks without sacrificing the richness of fused information. Our empirical results illustrate that effective privacy safeguards can coexist with accurate analytics, mitigating concerns that privacy protections necessarily degrade insights. We believe that the future direction of differential privacy research in marketing lies in studying the systemic impact of privacy-preserving methods on marketing functions and consumer behavior, exploring broader use cases and refining metrics for real-time, firm-level decision making. Beyond our methodological contribution, the ultimate goal is to ensure that data-driven innovations—now fundamental to many marketing practices—remain compatible with stringent data protection, thereby fostering trust and transparent engagement between firms and their customers.

Acknowledgments

The authors thank Fred Feinberg for his outstanding support in this project. Also, special thanks to Elizabeth E. Bruch, A. Yeşim Orhun, Nigel Melville, and Mengyao Huang for their valuable feedback on earlier versions of this paper. This work benefited from feedback from colleagues at Reichman University as well as feedback from participants of the University of Michigan, Tel Aviv University, and Emory University marketing, tech and law, and information systems seminars and the *Marketing Science* conference. Finally, the authors thank their colleagues at the partnering telecom carrier (trademark redacted due to the nondisclosure agreement) who supported their vision, provided invaluable expertise, ensured data quality, and went above and beyond to protect their customers' privacy. The authors received no payment or compensation of any form in conducting this research. The claims and views of this paper are solely those of the authors. The telecom carrier sharing the data used has rights to review and request changes to the paper for any material that may disclose its brand name, trademarks, and trade secrets. Although this project has benefited from all of this support, all mistakes are solely the authors' own. Longxiu Tian and Dana Turjeman contributed equally to the manuscript.

Appendix A. Comparison with Benchmarks

Our empirical results show that in the context of data fusion, the proposed framework ("global PPDF" or PPDF in short) yields lower mean absolute error (MAE) than local PPDF. Figure A.1 demonstrates the interplay between data reconstruction accuracy in terms of MAE in crossimputation ability and the percentage of respondents who are uniquely reidentified (see Section 5.1.2) across the telecom carrier's CRM and customer satisfaction survey. The non-DP model shown in red in Figure A.1, although providing the most accurate reconstruction, shows the highest percentage of reidentified respondents, underlining a pronounced trade-off between data utility and privacy.

The proposed (global) PPDF shown in black in Figure A.1 consistently shows zero respondents reidentified across all ε settings in both data sets. This illustrates its superior capacity to safeguard privacy, effectively preventing any unique reidentification of individuals, which is critical as identifying even a single person can lead to significant privacy breaches and potential misuse of personal data. Conversely, local PPDF, which is shown in blue in Figure A.1, displays a highly variable trade-off between accuracy and privacy. Although it reduces the number of reidentified respondents compared with the non-DP model, it does not achieve the complete indistinguishability seen under global DP across data sets and across ε levels. This variance highlights the nuanced balance between maintaining data utility and enhancing privacy, which local DP navigates differently from global DP.

Within each DP model, variations in ε levels further influence the privacy-accuracy dynamic. Lower ε values, signifying tighter privacy controls, are expected to reduce the number of reidentified individuals; however, this often results in increased MAE values indicative of less precise data reconstruction. Higher ε values, although improving reconstruction accuracy, typically allow for higher percentages of respondent reidentification across the two data sets; this demonstrates that these trends are generally consistent, but the specific impacts of model choice and ε levels can differ. Notably, under global DP, PPDF maintains zero reidentifications across all configurations of ε. Hence, compared with local DP, global DP performs better in both accuracy and privacy.

Beyond variations in ε levels, another element highly affecting the privacy risk is the number of common variables. To this end, we present a series of simulation studies in Appendix B and show how PPDF consistently ensures the desired privacy budget, even across varying dimensions of common variables between data sets.

Appendix B. Simulation Exercise

In Appendix A, we presented the sensitivity of various models to the choice of privacy guarantees ε. The next simulation serves as the vehicle to assess the sensitivity of the framework to (1) different levels of differential privacy, (2) local versus global differential privacy, and (3) the influence of the proportion of "common variables" relative to all variables on reidentification.

Two data sets, A and B, are synthetically created in each run containing both uncommon and common variables. We initialize the simulation with random keys to ensure reproducibility and independence during data generation. Three decoders, one for each set of uncommon and common variables, are pseudorandom initialized with Glorot Normal parameter weights. We choose a feed-forward neural network architecture for the simulation decoder that contains different activation functions than the proposed model's function to prevent perfect recoverability. Latent variables are sampled from a standard normal distribution and passed through the decoder functions to generate synthetic variables. These variables are then binarized to form the data sets. The binarization is based on a cutoff of greater than or less than zero.

Figure A.1. (Color online) Model Comparisons in Terms of the Reidentification Metric

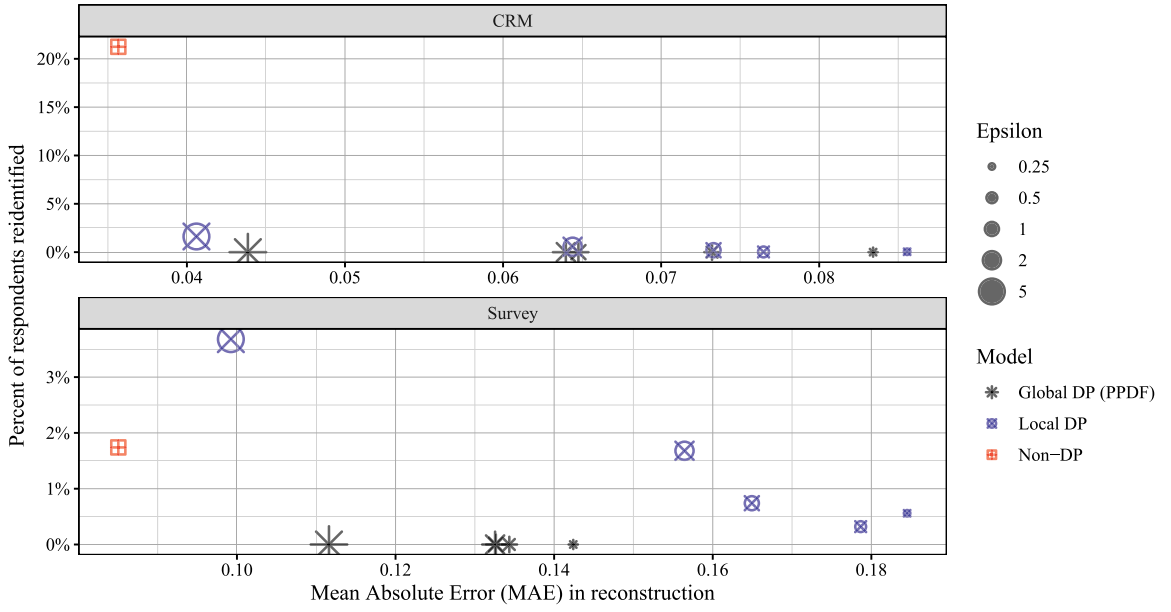
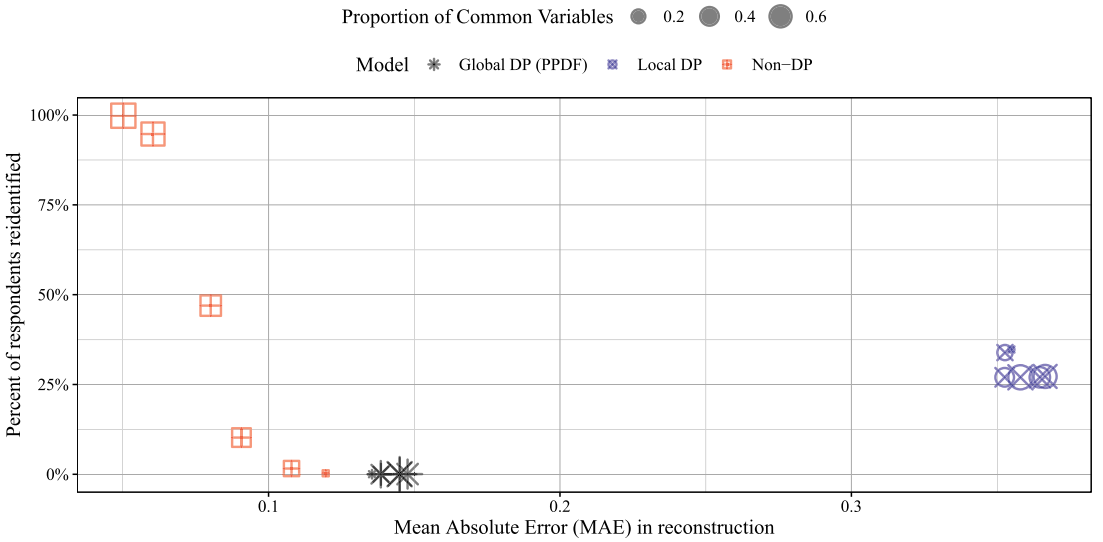


Figure B.1. (Color online) Simulation Results: Effect of the Proportions of Common Variables and Model on Accuracy and Reidentification Ability



Note. $\varepsilon = 1.00$ when with differential privacy.

If differential privacy is global, then the split data sets are estimated via differential privacy-stochastic variational inference. If differential privacy is local, then the data sets are processed via the generalized random response algorithm (Kairouz et al. 2016). Both local and global differential privacy take as input the differential privacy parameter ε. The simulated data and estimation were structured to mimic telecom carrier's data and inference parameters: 1,000 iterations, 1,000 posterior draws, minibatch size of 75, seven layers, 256-dimensional latent variables, and 2.5×10^{-4} learning rate for the optimizer. Estimation is made on training, validation, and test splits of, respectively, 40,000, 5,000, and 5,000 rows.

In Figure B.1, we present estimation results of the reconstruction of data set A (data set B's results are identical as they are symmetric) when $\varepsilon = 1.00$: the ability to reconstruct data set A from data set B following the data fusion process—decoding of one data set into another.

A critical observation from the graph is the distinctly higher reidentification risk associated with the non-DP model across varying proportions of common variables. Both non-DP and local DP consistently present higher percentages of respondents identified relative to global DP (i.e., PPDF), thus highlighting their vulnerability in safeguarding user privacy. However, in low proportions of common variables, non-DP is improving—having lower reidentification risk; this aligns with the work of Sweeney (1997) and may aid managers in deciding how to deconstruct data sets for future storage to prevent harsher results of data breaches. The local DP model exhibits higher reidentification risk and lower accuracy at all proportions of common variables.

We present the results of different levels of ε in Figure B.2. These results suggest inferior outcomes of local DP across various levels of privacy and proportions of common variables: overfitting the data (superior ability to self-impute data and inferior ability to crossimputation) (see Table B.1) and unable

Figure B.2. (Color online) Simulation Results: Effect of the Proportions of Common Variables, Model, and ε Levels on Accuracy and Reidentification Ability

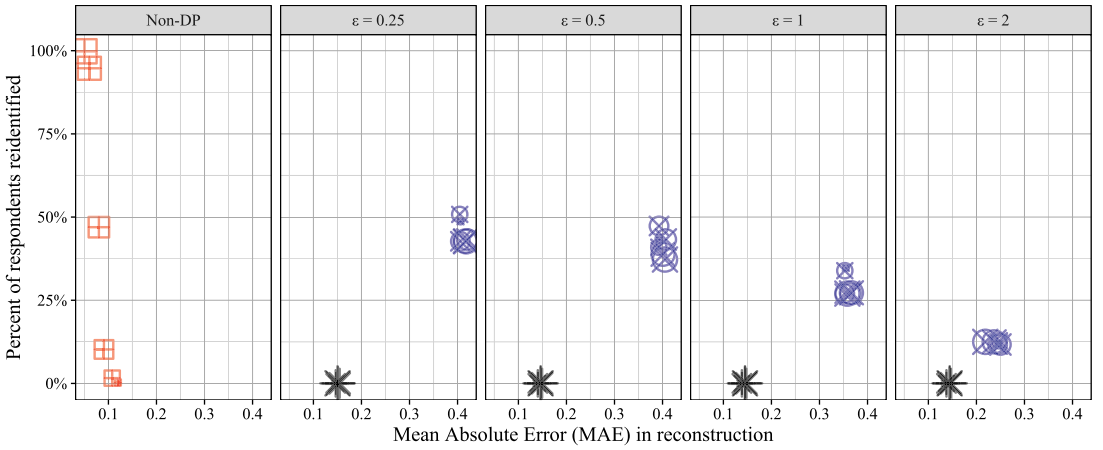


Table B.1. Simulations for Global, Local, and No Differential Privacy for 60% Common Variables Across Data Sets A and B

	Data set A			Data set B		
Model	MAE train	MAE test	MAE holdout	MAE train	MAE test	MAE holdout
Self-reconstruction						
Non-DP	0.04 (0.05)	0.05 (0.05)	0.05 (0.05)	0.04 (0.04)	0.04 (0.05)	0.04 (0.05)
DP ($\varepsilon = 0.25$)	0.16 (0.17)	0.16 (0.17)	0.16 (0.17)	0.15 (0.16)	0.15 (0.16)	0.15 (0.16)
Local DP ($\varepsilon = 0.25$)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)
DP ($\varepsilon = 0.5$)	0.16 (0.17)	0.16 (0.17)	0.16 (0.17)	0.15 (0.17)	0.15 (0.17)	0.15 (0.17)
Local DP ($\varepsilon = 0.5$)	0.03 (0.04)	0.03 (0.04)	0.04 (0.04)	0.03 (0.04)	0.03 (0.04)	0.03 (0.04)
DP ($\varepsilon = 1.0$)	0.16 (0.17)	0.16 (0.17)	0.16 (0.17)	0.14 (0.17)	0.14 (0.17)	0.14 (0.17)
Local DP ($\varepsilon = 1.0$)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.04 (0.05)	0.04 (0.05)	0.04 (0.05)
DP ($\varepsilon = 2.0$)	0.15 (0.17)	0.15 (0.17)	0.15 (0.17)	0.14 (0.17)	0.14 (0.17)	0.14 (0.17)
Local DP ($\varepsilon = 2.0$)	0.05 (0.06)	0.05 (0.06)	0.05 (0.06)	0.05 (0.06)	0.05 (0.06)	0.05 (0.06)
Crossimputation						
Non-DP	0.05 (0.07)	0.05 (0.07)	0.05 (0.07)	0.05 (0.07)	0.05 (0.07)	0.05 (0.07)
DP ($\varepsilon = 0.25$)	**0.15** (0.16)	**0.15** (0.16)	**0.15** (0.16)	**0.15** (0.16)	**0.15** (0.16)	**0.15** (0.16)
Local DP ($\varepsilon = 0.25$)	0.42 (0.17)	0.43 (0.18)	0.42 (0.17)	0.42 (0.17)	0.43 (0.18)	0.42 (0.17)
DP ($\varepsilon = 0.5$)	**0.15** (0.17)	**0.15** (0.17)	**0.15** (0.17)	**0.15** (0.17)	**0.15** (0.17)	**0.15** (0.17)
Local DP ($\varepsilon = 0.5$)	0.41 (0.17)	0.41 (0.17)	0.41 (0.17)	0.41 (0.17)	0.41 (0.17)	0.41 (0.17)
DP ($\varepsilon = 1.0$)	**0.14** (0.17)	**0.14** (0.17)	**0.14** (0.17)	**0.14** (0.17)	**0.14** (0.17)	**0.14** (0.17)
Local DP ($\varepsilon = 1.0$)	0.35 (0.15)	0.35 (0.15)	0.36 (0.15)	0.35 (0.15)	0.35 (0.15)	0.36 (0.15)
DP ($\varepsilon = 2.0$)	**0.14** (0.17)	**0.14** (0.17)	**0.14** (0.17)	**0.14** (0.17)	**0.14** (0.17)	**0.14** (0.17)
Local DP ($\varepsilon = 2.0$)	0.22 (0.10)	0.22 (0.10)	0.22 (0.10)	0.22 (0.10)	0.22 (0.10)	0.22 (0.10)

Notes. Larger ε means less privacy and more accurate results. Posterior standard errors are given in parentheses.

to accurately construct and protect respondents of the simulated data. Global DP (PPDF), on the other hand, is consistent in both having relatively high accuracy and having no reidentification risk of even a single individual.

Appendix C. Model Comparisons for Predicted Targeting Analysis

Table C.1 compares a range of models under the predicted targeting scenario using the area under the receiver operating characteristic curve (AUC-ROC) metric for both training and holdout sets. The first block in Table C.1 features the full model without differential privacy (PPDF no DP), which attains the highest holdout AUC-ROC (0.948). XGBoost is a non-Bayesian reference benchmark, and accordingly, no posterior intervals are reported for it in Table C.1. The next rows in Table C.1 capture various baseline methods (e.g., "COMP no flow" and "COMP SUR"), each omitting certain modeling innovations, like normalizing flows or crossencoder components. These methods generally perform worse on the holdout set, highlighting the importance of joint generative modeling in PPDF.

In the lower block in Table C.1, the differentially private versions of PPDF demonstrate how privacy constraints affect predictive performance; the more stringent the privacy budget ε is, the lower the AUC-ROC on the holdout set is. Notably, DP with $\varepsilon = 5.0$ yields a holdout AUC-ROC of 0.789—above more restrictive privacy levels ($\varepsilon \leq 2.0$) but below the nonprivate PPDF. These results align with the discussion in the main text, showing that although DP inevitably reduces accuracy, larger ε-values can strike a better

Table C.1. Model Performance Comparison in Terms of Area Under the Receiver Operating Characteristic Curve for Train and Holdout Data Sets

AUC-ROC		Train				Holdout			
Model	Specification	Mean	St. dev.	1% HPD	99% HPD	Mean	St. dev.	1% HPD	99% HPD
Full model, non-DP	PPDF (no DP)	0.979	0.001	0.977	0.981	0.948	0.004	0.939	0.956
Comparison specifications, non-DP	XGBoost (non-Bayesian)	0.971	N/A	N/A	N/A	0.857	N/A	N/A	N/A
Comparison specifications, non-DP	No Normalizing Flow	0.913	0.002	0.908	0.918	0.736	0.007	0.722	0.751
Comparison specifications, non-DP	SUR-NN	0.931	0.002	0.926	0.936	0.735	0.008	0.718	0.752
Comparison specifications, non-DP	SUR	0.926	0.002	0.921	0.930	0.734	0.007	0.718	0.750
Comparison specifications, non-DP	BCCA	0.925	0.002	0.919	0.930	0.732	0.008	0.712	0.749
Full model, DP	DP ($\varepsilon = 5.0$)	0.944	0.002	0.940	0.948	0.789	0.008	0.771	0.807
Full model, DP	DP ($\varepsilon = 2.0$)	0.952	0.003	0.945	0.958	0.623	0.016	0.587	0.659
Full model, DP	DP ($\varepsilon = 1.0$)	0.958	0.003	0.952	0.964	0.555	0.016	0.518	0.592
Full model, DP	DP ($\varepsilon = 0.5$)	0.959	0.003	0.953	0.966	0.548	0.016	0.509	0.587
Full model, DP	DP ($\varepsilon = 0.25$)	0.961	0.003	0.955	0.968	0.515	0.016	0.477	0.552

Notes. Highest posterior density intervals are generated from 1,000 posterior draws except for XGBoost. BCCA, Bayesian canonical correlation analysis; N/A, not applicable; St. dev., standard deviation; SUR, seemingly unrelated regression; SUR-NN, seemingly unrelated regression (neural network).

balance between protecting users' anonymity and retaining predictive power. The 1% and 99% highest posterior density intervals come from 1,000 posterior draws in the Bayesian models, reflecting the sampling variability in each method's performance estimates.

Appendix D. Sensitivity Analysis: Distributional Overlap Assumption

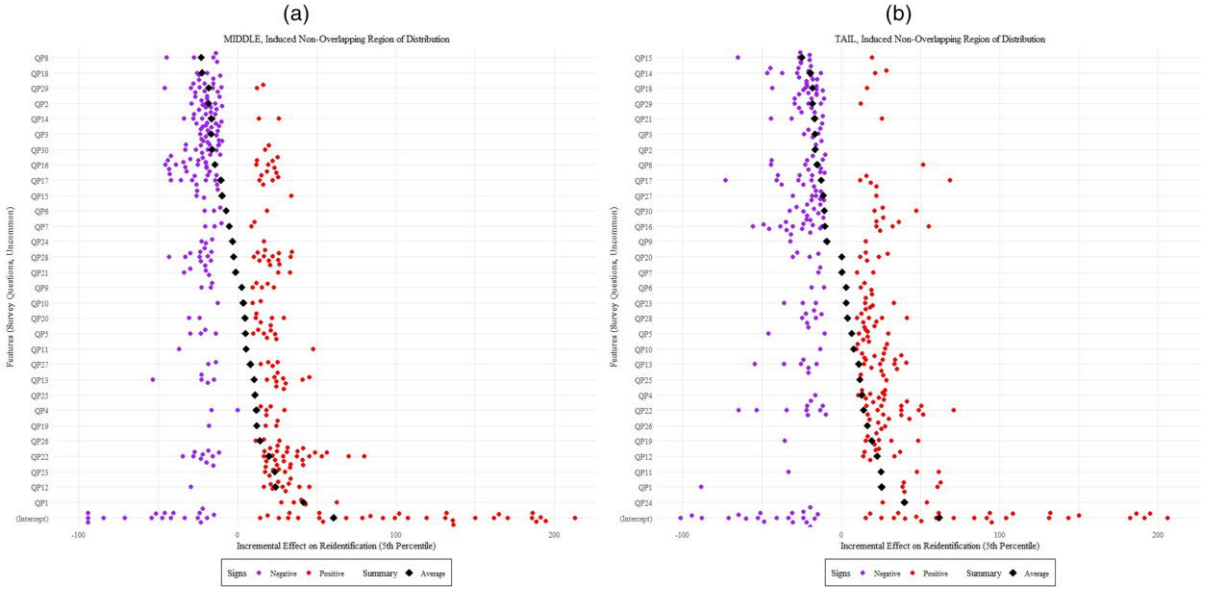
This analysis addresses how the degree of distributional overlap affects the generalizability of our results. A unique feature of our two data sets is that the survey respondents (whether from external or internal validation surveys) are actual customers present in the CRM, implying an assumption—and in this case, a reality—of distributional overlap. In other words, the CRM can be seen as a superset of the surveyed group. However, in many data fusion contexts, one cannot guarantee that the same population truly underlies both sources. Consequently, we simulate overlap violations on our focal data and examine the effects on reidentifiability.

We remove individuals from the CRM database if they hold certain corresponding variable values among the 68 common variables in the training data. Specifically, in one scenario, we remove those whose values lie in the middle 20% of the distribution for a given variable; in another, we remove those in the two tails of the distribution (10% each). The holdout validation data are kept intact. Each such removal yields a distinct "overlap-violation scenario," and we retrain PPDF from scratch, evaluating the posterior predictive reidentifiability for individuals in the holdout validation survey. We then compute any incremental change in the fifth-percentile reidentifiability metric for each holdout respondent, comparing the base PPDF ($\varepsilon = 5.00$) with each overlap-violation scenario. Finally, we segment users according to which variables made them more or less reidentifiable under these induced violations.

Prior to the analysis, we did not have a prespecified hypothesis about whether artificially violating overlap would help or hinder reidentification or about the directionality of any particular variable's effect. Reidentifiability, after all, is relative; we can imagine that removing midrange observations for some variables might simplify distributions, thus making true outliers stand out more easily. Conversely, removing outliers in the tails can induce greater shrinkage in crossimputed values, which may reduce distinctiveness among individuals. Moreover, these two effects can coexist. One might see, for instance, a global effect in which overlap violations reduce reidentifiability on average but a local or segment-specific effect that moves in the opposite direction for a subset of the population.

Figure D.1 displays two bee swarm plots summarizing results from the resulting 136 runs of overlap violations induced in the middle and in the tails of the common-variable distributions. Panels (a) and (b) in Figure D.1 aggregate the incremental changes observed in the fifth-percentile reidentifiability metric across scenarios, with one scenario per common variable. The *y* axis in Figure D.1 lists these variables sorted by the average directionality (black diamonds) of the incremental change in reidentifiability. At the bottom of panels (a) and (b) in Figure D.1 is the label "intercept." We see that overlap violations often lead to a positive incremental shift in reidentification. Interpreted another way, retaining fuller overlap can make it harder to reidentify individuals. Still, this overall trend masks heterogeneity; certain variables, such as QP24 under tail removal or QP1 under midremoval, present more uniform patterns of higher reidentification likelihood whenever the variable overlaps are removed.

Figure D.1. (Color online) Illustrative Plots of the Incremental Effect on Reidentification Displayed at the Fifth Percentile



Notes. Negative effects (purple dots) and positive effects (red dots) indicate how certain survey responses influence the likelihood of being reidentified (*uncommon* survey variables). The black diamonds mark average or summary estimates for each question. Panels (a) and (b) highlight the differences in effect magnitudes at the "middle" vs. "tail" portions of the distribution. (a) Reidentification sensitivity (middle of distribution). (b) Reidentification sensitivity (tail of distribution).

Taken together, these findings reveal that overlap violations typically reduce reidentifiability in a global sense, yet the segment-specific impacts vary widely. In some instances, the tail or middle removal yields subgroups whose outlier characteristics lead to easier identification; in others, forced shrinkage in the training data reduces the ability of the fused model to distinguish individuals. Although a deeper exploration of the interplay between each common variable's distribution and privacy outcomes lies beyond our current scope, we view this as a promising direction for future work. Such an investigation could have broader implications for how firms manage data releases or how public institutions balance data transparency against privacy risks in large-scale settings, like data clean rooms or multimodal research.

Endnotes

[1] As a matter of nomenclature, we avoid the term "identify" in the context of privacy to avoid confusing it with its common usage in econometrics and statistics, and instead, we refer to "reidentify," which is defined as linking anonymous data records to additional data sources where PII exists.

[2] The nouns user, individual, person, customer, and consumer will be used interchangeably to describe people whose information, some of which may be private and/or identifiable, is held by companies.

[3] Here, we do not take into consideration the sensitivity of the data set or the material risks associated with being revealed. Instead, we consider the legal and managerial aims of keeping all customers unidentified.

[4] Privacy budget refers to a core concept in differential privacy and privacy-enhancing technologies; it is a quantitative allowance (often denoted ε) that tracks how much privacy loss can be tolerated, and it is operationalized as how much noise must be added to obfuscate individuals while still retaining useful analytic signals. Smaller budgets imply stronger privacy but less accuracy, whereas larger budgets imply weaker privacy but more accurate outputs. We discuss this further in Section 4.2.

[5] Regarding confidentiality, the data were provided for this research under a nondisclosure agreement. Some figures and identifiers of the telecom carrier and of their customers are removed or obscured. The percentage range is with respect to total customers across all three major U.S. telecom carriers at the time of writing.

[6] A concern of the carrier is the accuracy of the common variables in the survey as they are self-reported. We leveraged the internal validation survey's ground-truth linkage and found only trivial differences. This is unsurprising as the survey questions are designed to ensure accurate recall (i.e., asking for income brackets rather than a specific value).

[7] Source code can be found in the Online Appendix.

[8] An assumption made here is that the TTM of the past year will carry on to the next year.

[9] XGBoost serves as a proxy for the firm's existing nonprivatized, nondata fusion workflow for churn prediction. Note that whether fitting XGBoost using samples from the so-called "joint" distribution or directly on $\mathbf{x}_i^{(\mathrm{CRM})}$, the only actual input data in both cases are the $\mathbf{x}_i^{(\mathrm{CRM})}$.

[10] More details on the workflow, robustness checks, and hyperparameter tuning are available in Online Appendix EC.9.

References

Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. *Proc. 2016 ACM SIGSAC Conf. Comput. Comm. Security* (Association for Computing Machinery, New York), 308–318.

Anand P, Lee C (2023) Using deep learning to overcome privacy and scalability issues in customer data transfer. *Marketing Sci.* 42(1):189–207.

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *J. Marketing Res.* 55(1):80–98.

Bradburn NM, Sudman S, Blair E, Locander W, Miles C, Singer E, Stocking C (1979) *Improving Interview Method and Questionnaire Design: Response Effects to Threatening Questions in Survey Research* (Jossey-Bass, San Francisco).

Bradlow ET, Zaslavsky AM (1999) A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *J. Amer. Statist. Assoc.* 94(445):43–52.

Cai C, Sang Y, Tian H (2022) A multimodal differential privacy framework based on fusion representation learning. *Connection Sci.* 34(1):2219–2239.

Carey C, Dick T, Epasto A, Javanmard A, Karlin J, Kumar S, Muñoz Medina A, et al. (2023) Measuring re-identification risk. *Proc. ACM Management Data* 1(2):149.

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 785–794.

Chen X, Simchi-Levi D, Wang Y (2022) Privacy-preserving dynamic personalized pricing with demand learning. *Management Sci.* 68(7):4878–4898.

Dankar FK, El Emam K, Neisa A, Roffey T (2012) Estimating the re-identification risk of clinical data sets. *BMC Medical Informatics Decision Making* 12(1):66.

De Haan E, Verhoef PC, Wiesel T (2015) The predictive ability of different customer feedback metrics for retention. *Internat. J. Res. Marketing* 32(2):195–206.

Dehghan M, Khosravian E, Golfar Z, Shahbazi H (2022) P2DF: Privacy-preserving data fusion protocol. *2022 12th Internat. Conf. Comput. Knowledge Engrg. (ICCKE)* (IEEE, Piscataway, NJ), 211–218.

Ding W, Jing X, Yan Z, Yang LT (2019) A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion. *Inform. Fusion* 51:129–144.

Dunn HL (1946) Record linkage. *Amer. J. Public Health Nations Health* 36(12):1412–1416.

Dwork C, McSherry F, Nissim K, Smith A (2006a) Calibrating noise to sensitivity in private data analysis. Halevi S, Rabin T, eds. *Theory of Cryptography. TCC 2006*, Lecture Notes in Computer Science, vol. 3876 (Springer, Berlin), 265–284.

Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M (2006b) Our data, ourselves: Privacy via distributed noise generation. *Annual Internat. Conf. Theory Appl. Cryptographic Techniques* (Springer, Berlin, Heidelberg), 486–503.

Evans G, King G, Schwenzfeier M, Thakurta A (2023) Statistically valid inferences from privacy-protected data. *Amer. Political Sci. Rev.* 117(4):1275–1290.

Feit EM, Beltramo MA, Feinberg FM (2010) Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Sci.* 56(5):785–800.

Feit EM, Wang P, Bradlow ET, Fader PS (2013) Fusing aggregate and disaggregate data with an application to multiplatform media consumption. *J. Marketing Res.* 50(3):348–364.

Gati NJ, Yang LT, Feng J, Nie X, Ren Z, Tarus SK (2021) Differentially private data fusion and deep learning framework for cyber–physical–social systems: State-of-the-art and perspectives. *Inform. Fusion* 76:298–314.

Gilula Z, McCulloch RE, Rossi PE (2006) A direct approach to data fusion. *J. Marketing Res.* 43(1):73–83.

Hassani S, Dackermann U, Mousavi M, Li J (2024) A systematic review of data fusion techniques for optimized structural health monitoring. *Inform. Fusion* 103:102136.

Huang Q, Zhang J, Zeng Z, He D, Ye X, Chen Y (2023) PPDF-fedTMI: A federated learning-based transport mode inference model with privacy-preserving data fusion. *Simulation Model. Practice Theory* 129:102845.

Kairouz P, Bonawitz K, Ramage D (2016) Discrete distribution estimation under local privacy. *Internat. Conf. Machine Learn*. (PMLR, New York), 2436–2444.

Kaissis G, Ziller A, Passerat-Palmbach J, Ryffel T, Usynin D, Trask A, Lima I Jr, et al. (2021) End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence* 3(6):473–484.

Kasiviswanathan SP, Lee HK, Nissim K, Raskhodnikova S, Smith A (2011) What can we learn privately? *SIAM J. Comput.* 40(3):793–826.

Kim K, Tanuwidjaja HC (2021) *Privacy-Preserving Deep Learning: A Comprehensive Survey* (Springer, Singapore).

Kingma DP, Welling M (2019) An introduction to variational autoencoders. *Foundations Trends Machine Learn.* 12(4):307–392.

Klymenko A, Meisenbacher S, Lilova I, Matthes F (2024) Investigating the motivational factors influencing managerial decisions to adopt privacy-enhancing technologies. *ECIS 2024 Proc.* (AIS eLibrary).

Korganbekova M (2023) Balancing user privacy and personalization. Work in progress, Kellogg School of Management, Northwestern University, Evanston, IL.

Lemmens A, Gupta S (2020) Managing churn to maximize profits. *Marketing Sci.* 39(5):956–973.

Li S, Schneider MJ, Yu Y, Gupta S (2023) Reidentification risk in panel data: Protecting for *k*-anonymity. *Inform. Systems Res.* 34(3):1066–1088.

Lin T, Misra S (2022) Frontiers: The identity fragmentation bias. *Marketing Sci.* 41(3):433–440.

Liu J, Lou J, Liu J, Xiong L, Pei J, Sun J (2021) Dealer: An end-to-end model marketplace with differential privacy. *Proc. VLDB Endowment* 14(6):957–969.

Lobschat L, Mueller B, Eggers F, Brandimarte L, Diefenbach S, Kroschke M, Wirtz J (2021) Corporate digital responsibility. *J. Bus. Res.* 122:875–888.

Malshe A, Colicev A, Mittal V (2020) How main street drives wall street: Customer (dis)satisfaction, short sellers, and abnormal returns. *J. Marketing Res.* 57(6):1055–1075.

McCarthy DM, Oblander ES (2021) Scalable data fusion with selection correction: An application to customer base analysis. *Marketing Sci.* 40(3):459–480.

Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. *2008 IEEE Sympos. Security Privacy (SP 2008)* (IEEE, Piscataway, NJ), 111–125.

Neumann N, Tucker CE, Kaplan L, Mislove A, Sapiezynski P (2024) Data deserts and black boxes: The impact of socio-economic status on consumer profiling. *Management Sci.* 70(11):8003–8029.

Prediger L, Loppi N, Kaski S, Honkela A (2022) d3p—A Python package for differentially-private probabilistic programming. *Proc. Privacy Enhancing Tech.* 2022(2):407–425.

Rafieian O, Yoganarasimhan H (2021) Targeting and privacy in mobile advertising. *Marketing Sci.* 40(2):193–218.

Rezende D, Mohamed S (2015) Variational inference with normalizing flows. *Internat. Conf. Machine Learn*. (PMLR, New York), 1530–1538.

Rogers R, Subramaniam S, Peng S, Durfee D, Lee S, Kancha SK, Sahay S, Ahammad P (2021) LinkedIn's audience engagements API: A privacy preserving data analytics system at scale. *J. Privacy Confidentiality* 11(3).

Ruggles S (2024) When privacy protection goes wrong: How and why the 2020 Census confidentiality program failed. *J. Econom. Perspect.* 38(2):201–226.

Ruohonen J, Hjerppe K (2022) The GDPR enforcement fines at glance. *Inform. Systems* 106:101876.

Schoenmueller V, Netzer O, Stahl F (2020) The polarity of online reviews: Prevalence, drivers and implications. *J. Marketing Res.* 57(5):853–877.

Shelake VM, Shekokar N (2017) A survey of privacy preserving data integration. *2017 Internat. Conf. Electr. Electronics Comm. Comput. Optim. Techniques (ICEECCOT)* (IEEE, Piscataway, NJ), 59–70.

Sisodia A, Burnap A, Kumar V (2025) Generative interpretable visual design: Using disentanglement for visual conjoint analysis. *J. Marketing Res.* 62(3):405–428.

Steinke T (2022) Composition of differential privacy & privacy amplification by subsampling. Preprint, submitted October 26, https://arxiv.org/pdf/2210.00597.

Swait J, Andrews RL (2003) Enriching scanner panel models with choice experiments. *Marketing Sci.* 22(4):442–460.

Sweeney L (1997) Weaving technology and policy together to maintain confidentiality. *J. Law Medicine Ethics* 25(2–3):98–110.

Sweeney L (2000) Simple demographics often identify people uniquely. *Health (San Francisco)* 671(2000):1–34.

Sweeney L (2002) k-anonymity: A model for protecting privacy. *Internat. J. Uncertainty Fuzziness Knowledge-Based Systems* 10(05):557–570.

Takagi S, Takahashi T, Cao Y, Yoshikawa M (2020) P3GM: Private high-dimensional data release via privacy preserving phased generative model. *2021 IEEE 37th Internat. Conf. Data Engineering (ICDE)* (IEEE, Piscataway, NJ), 169–180.

Tian Z, Dew R, Iyengar R (2024) Mega or micro? Influencer selection using follower elasticity. *J. Marketing Res.* 61(3):472–495.

Turjeman D, Feinberg FM (2024) When the data are out: Measuring behavioral changes following a data breach. *Marketing Sci.* 43(2):440–461.

Unger M, Shapira B, Rokach L, Livne A (2018) Inferring contextual preferences using deep encoder-decoder learners. *New Rev. Hypermedia Multimedia* 24(3):262–290.

U.S. Census Bureau (2021) Disclosure avoidance for the 2020 Census: An introduction. Report, U.S. Census Bureau, Washington, DC.

Verhoef PC (2003) Understanding the effect of customer relationship management efforts on customer retention and customer share development. *J. Marketing* 67(4):30–45.

Wang Q, Yang K (2024) Privacy-preserving data fusion for traffic state estimation: A vertical federated learning approach. *Transportation Sci.* 168 (104743).

Longxiu Tian is an assistant professor of marketing at the University of North Carolina Kenan–Flagler Business School. His research is in developing Bayesian econometric and probabilistic machine learning models for measuring marketing effectiveness and incrementality in customer relationship management. He received his PhD in marketing and scientific computing from the University of Michigan.

Dana Turjeman is an assistant professor of Marketing and Data Science at the Arison School of Business, Reichman University. She studies decision making in digital environments, with a focus on privacy, uncertainty, and platform behavior. She develops causal and behavioral models to study privacy risks, privacy-preserving data integration, and financial decision-making on digital platforms. She earned her PhD in Quantitative Marketing from the University of Michigan.

Samuel Levy is an assistant professor of business administration (marketing) at the University of Virginia Darden School of Business. His research focuses on customer value management, targeting, and personalization. He develops Bayesian econometric and probabilistic machine learning methods to improve marketing decision making and model consumer choice. He earned his PhD in marketing from the Carnegie Mellon University Tepper School of Business.