Data-Sharing Economy: Value-Addition from Data meets Privacy

Piyush Bagad*
IIT Kanpur & Adobe

Subrata Mitra
Adobe

Sunny Dhamnani*
Adobe

Atanu R. Sinha
Adobe

Raunak Gautam*
IIT Delhi & Adobe

Haresh Khanna*
IIT Roorkee & Adobe

ABSTRACT

The need for improved segmentation, targeting, personalization fuel the practice of data sharing among companies. Concurrently, data sharing faces the headwind of new laws emphasizing users’ privacy in data. Under the premise that sharing of data occurs from a provider to a recipient, we propose a practicable approach of generating representational data for sharing that achieves value-addition for the recipient’s tasks while preserving privacy of users. Prior art shows that the mechanism to improve value-addition inevitably weakens privacy in the generated data. In a first of a kind contribution, our system offers tunable controls to adjust the extent of privacy desired by the provider and the extent of value-addition expected by the recipient. Our experiments on a public data show that under common organizational practice of data-sharing, data generation for value-addition is achievable while preserving privacy. Our demonstration starkly shows the trade-off between privacy-protection and value addition, through user-controlled knobs and offers a prototype of a platform for data sharing which is mindful of this trade-off.

CCS CONCEPTS

• **Security and privacy → Privacy protections**; *Social aspects of security and privacy*; *Usability in security and privacy*.

KEYWORDS

Data sharing, Privacy, Utility, Generative adversarial networks

ACM Reference Format:
Piyush Bagad, Subrata Mitra, Sunny Dhamnani, Atanu R. Sinha, Raunak Gautam, and Haresh Khanna. 2021. Data-Sharing Economy: Value-Addition from Data meets Privacy. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM ’21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437963.3441712

1 INTRODUCTION

Two major forces are at work in the data-sharing economy. The first force, data-sharing among companies, is accelerating. This is driven by the need for improvement in *tasks* such as segmentation, targeting and personalization to achieve higher relevance with users and customers (hereafter, users) [2, 5]. Any single company *lacks* data about *all touch points* (e.g. a user’s interactions with website and an app) and *all profiling* characteristics of users. Data sharing fills this gap by having more complete data on users, enabling companies to obtain value-addition by building more accurate models of user behavior for these tasks. The second force, user-privacy in data (hereafter, privacy), is gathering momentum as seen in laws such as GDPR (General Data Protection Regulation) in Europe and CCPA (California Consumer Privacy Act). These two forces tend to collide [4, 8]. For example, obfuscating data for sharing purposes can achieve privacy, but sacrifice value to the recipient when important attributes are obfuscated, reducing effectiveness for these tasks. The provider’s goals are maintaining users’ privacy in the shared data and meeting expectations of recipient to perform value-added modeling. Models for value addition include clustering and regression (classification) analysis. These popular analyses are built on dependency structure among variables (attributes). On the other hand, when attributes have high dependency among them, knowing some attributes allow inference about others even in their absence. Traditional techniques of privacy preservation such as anonymization [10] and differential privacy [1] rely on adding noise to individual attributes and thus, may lose the dependency structure which in turn degrades value-addition. In anonymization (particularly k-anonymization [10]), data are modified such that each user’s information is indistinguishable from at least k-1 other users. However, k-anonymity is prone to re-identification attacks through auxiliary knowledge acquired by the attacker [6]. In differential privacy, the data access (release) mechanism is modified by introducing controlled noise so that presence or absence of any record cannot be inferred by observing the output. However, it is empirically noted [1, 14] that for high-dimensional data, differential privacy is either computationally expensive or adds too much perturbation leaving the transformed data less useful for value-addition. Moreover, these techniques may not provide safeguard against leakage of information that occurs through combination of attributes.

To overcome some of these deficiencies, recent works propose privacy protection by sharing only *representational data*, instead of the original data. To generate representation from the original data, techniques such as Generative Adversarial Networks (GANs) [7, 12], Variational Autoencoders (VAEs) [12], Gaussian copula, etc. can be used. In general GANs and VAEs can preserve the statistical properties of the original data significantly better than other traditional statistical methods [13], thereby satisfying the need for generated data to provide significant value-addition for the recipient. A recent GAN-based approach [7] improves value-addition by including loss in first-order and second-order statistics and loss in classification accuracy for an ML task. However, the resultant loss of privacy is not explicitly modeled. Generated data is also susceptible to sophisticated privacy attacks if the generative model learns to mimic

[*] Currently, PB is at Wadhwani AI, SD at Gatech, RG at Apple & HK at Microsoft.

WSDM ’21, March 8–12, 2021, Virtual Event, Israel
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8297-7/21/03.
https://doi.org/10.1145/3437963.3441712

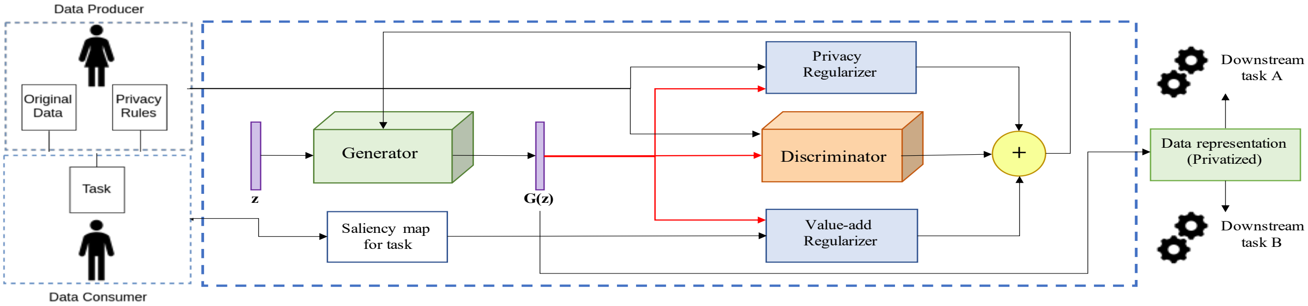

Figure 1: Work-flow of the solution: We propose GAN-based approach with dual regularisation of privacy and value-addition.

the original data too well, as may be the case for VAEs and GANs. Although, GANs might be better than VAEs for privacy preservation [12] because, unlike VAE, the generator in GAN does not have access to real data during the entire training process.

Even though at a high-level privacy is at odds with value-addition, in practice, privacy is often sought for some attributes in the data that do not fully overlap with the most important attributes for the recipient's intended tasks. The degree of privacy demanded among attributes can also vary. We argue that by having finer, attribute-wise control for privacy, a sweet-spot can be achieved where the shared-data can still provide desired privacy without sacrificing significant reduction in value-addition. Prior works [7, 12] that recognize both value-addition and privacy protection, do not consider this perspective and do not provide any control to navigate this space. In this paper, we propose a novel approach that uses generative models to offer attribute-wise *tunable controls* to adjust the extent of privacy desired by the provider *and* the extent of value-addition expected by the recipient. We assume that value-addition is judged by a *benchmark of tasks* (e.g. clustering, classification etc.) that is known to the provider. Provider also knows the attributes that are privacy-sensitive for the users. Using these information, we propose a novel GAN-based workflow (Figure 1) that employs two distinct regularizers corresponding to the dual goals of controlling privacy and preserving value-addition, for generating representational data to be shared with the recipient. Our approach can use tunable weights per attribute to provide granular control of the generated data while preserving other high-level statistical characteristics.

We make three contributions: (1) We make a case for explicit and attribute-wise granular control for balancing privacy and value-addition in data-sharing using generated representational data. (2) We propose a novel dual regularizer-based GAN workflow to achieve finer control between privacy and value-addition. (3) We empirically demonstrate that our approach can provide improved value-addition without compromising privacy while maintaining high-level statistical characteristics in the generated data.

2 METHODOLOGY

We assume the original data are tabular, denoted as $\mathbb{D} := \{\mathbf{x}_n\}_{n=1}^{N} \subset \mathbb{R}^{N \times F}$ where $\mathbf{x}_n = (x_{a_1}, x_{a_2}, ..., x_{a_F}) \in \mathbb{R}^F$ denotes a row corresponding to user n, $\forall n$ and $\mathbf{a}_j \in \mathbb{R}^N, \forall j$ denotes the attributes as columns. Given $\mathbb{D}$, we seek a function f such that the shareable representation $\mathbb{D}' = f(\mathbb{D})$ satisfies the desirable properties of privacy and value-addition. In privacy literature, attribute is one of four types: *identifier* (e.g., ID); *quasi-identifier* (QID) that identifies a user up to a larger group (e.g. occupation); *sensitive* that contains private attributes like disease status, race; and *others*, for the remaining attributes. We focus on *sensitive* and *QID* attributes for privacy preservation. To address value-addition, we use a *target* attribute, as a target label in a supervised setting. First, our notion of privacy is based on prevention of re-identification attack on the shared data so that sensitive attributes are not disclosed. Second, value addition is judged with regard to specific tasks at hand, such as classification and clustering. Third, we add regularizers for privacy and value-addition to the objective function of GAN, to generate representational data which can be shared. We extend Wasserstein GAN with Gradient Penalty (WGAN-GP)[3], which preserves well first and second order statistics for tabular data generation, to include regularizers explicitly for value-addition and privacy.

	None	Privacy
Precision	0.63	0.50
Recall	0.66	0.44
F1-score	0.63	0.46

Constraint vs classification metrics

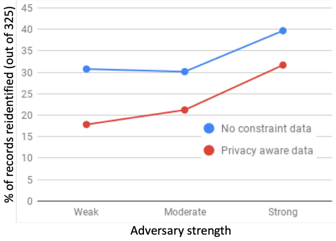

Figure 2: Evaluation of Privacy: (L) k-NN classification results: Observe that precision and recall drop for k-NN classification on privacy-aware generated data relative to that of unconstrained generation. (R) Re-identification attack: Observe that, relative to unconstrained data generation, for privacy-aware generated data, the number of successful re-identifications (1) is less, and (2) increases with adversary strength.

2.1 Privacy

Formally, privacy protection in terms of minimizing disclosure threats can be stated as $\min P\{d(r, A(\mathbb{D}', aux(r))) \leq \epsilon\}, \forall r \in \mathbb{D}$. Here, r is a record in $\mathbb{D}$, $aux(r)$ is the auxiliary information the attacker has about r and A is the adversary function to disclose information about users. Distance d is defined on the quasi-identifier attribute space instead of the entire attribute space. The privacy regularizer penalizes the generator for producing points that are too close to original data points in the space of attributes that can be potentially linked to re-identify individuals or infer information about their sensitive attributes. The privacy-aware objective is:

$$\min_{\theta_g}\max_{\theta_d}\; D(x)-D(G(z))+\lambda(\|\nabla_{\hat{x}}(D(\hat{x}))\|_2-1)^2-\lambda_p\mathbb{E}[d_p(G(z),\mathbb{D})]$$

Here, z denotes the noise $\sim \mathcal{N}(0,1)$ as input to generator G, D denotes discriminator in GAN, λ is the hyper-parameter controlling gradient-penalty regularizer [3] and λ_p is a hyper-parameter capturing effect of privacy regularizer. Distance of a generated point from k nearest original points is considered to assign penalty and d_p is exponential of weighted distance based on privacy weights as input.

$$d_p(G(z),\mathbb{D})=\exp\Big(-\sum_{i=1}^{k}\|G(z)-x_i\|_2^{\text{weighted}}\Big) \tag{1}$$

2.2 Value-addition

Consider that the recipient wants to do a binary classification task. The value-addition objective is to perform this task, by using $\mathbb{D}'$ in place of $\mathbb{D}$, without significant degradation in performance. That is, min $d'(g^*(\mathbb{D}),g^*(\mathbb{D}')$

$$\text{where } g^*(\mathbb{D}):=\operatorname*{argmin}_{g}\mathbb{E}_{(x,y)\in\mathbb{D}}[\mathcal{L}(g(x),y)]$$

Here, $g^*(\mathbb{D})$ is a model learned from data $\mathbb{D}$, $\mathcal{L}$ is the loss function for optimization and, d' is a distance metric between models, such as, accuracy. For the binary classification task, we extract the most *important* attributes through *saliency maps.* The generator is informed of the task by rewarding it for producing points close to real points in the space of these salient attributes. The objective that ensures value-addition is stated as:

$$\min_{\theta_g}\max_{\theta_d}\; D(x)-D(G(z))+\lambda(\|\nabla_{\hat{x}}(D(\hat{x}))\|_2-1)^2+\lambda_u\mathbb{E}[d_u(G(z),\mathbb{D})]$$

Here, λ_u is a hyper-parameter to tune the effect of the value-addition regularizer. For this regularizer, we consider weighted distance of a generated point $G(z)$ from k nearest original points. Distance d_u is crafted to apply opposing effect to that applied by equation 1. In other words, we want the generator to be rewarded if the importance-weighted distance between $G(z)$ and original points is less.

$$d_u(G(z),\mathbb{D})=1-\exp\Big(-\sum_{i=1}^{k}\|G(z)-x_i\|_2^{\text{weighted}}\Big) \tag{2}$$

3 EXPERIMENTS AND RESULTS

We use Adult Census public dataset [11], containing 6 real-valued, 8 categorical attributes, and 45,000 records. Attributes include race, income, etc. Sensitive attributes like race are to be privacy preserved. The value-added task is prediction of binary target label, whether annual income is greater than 50k USD. In a pre-processing step, real-valued attributes are normalized and categorical ones treated in two alternative ways: Gumbel soft-max and Auto-encoder.

Weights tunable by data provider: The data provider specifies the intensity of privacy required for each quasi-identifier or a combination of quasi-identifiers; intensity lies in [0,1]. These intensity scores are *privacy weights.* For the value-addition, we first extract the most important attributes through saliency maps. We assign weights in [0,1] to attributes in order of increasing saliency scores, referred as *importance weights.* Privacy and importance weights and the pre-processed data are fed to the generative model. The generator G and discriminator D train alternately, each trying to achieve self objective in competition with the other's objective.

Quality of generated data: First, the tSNE plot in Figure 3 (left) shows that the distribution of generated points is similar to the original data points. Second, a comparison of dependency structure of attributes between original data and our privacy-aware generated data is shown through a correlation matrix heat-map in Figure 3 (center). Note, privacy weights are applied to attributes 1 to 6 (marked by blue box). The correlation structure of attributes 1 to 6 is sufficiently different between original and generated data, while correlation structure for remaining attributes is reasonably preserved. Third, Frechet Inception Distance (FID) (Figure 3, right), along with G and D losses (not shown) stabilizes after 5000 iterations indicating convergence. FID is a statistical measure comparing first two order statistics of original and generated data and lower score means more closeness between generated and original data.

Experimentation for Privacy: Privacy protection is evaluated against *re-identification* and *attribute disclosure* attacks.

Re-identification Attack Model: We devise an attack to re-identify users from the shared data and some auxiliary information that the attacker has from other sources [6]. Formally, the attack model goes as follows. We call a subset of records T of the original data $\mathbb{D}$ as the *target* records that the attacker wants to re-identify. For each $r \in T$, we assume that the attacker has some auxiliary information $aux(r)$. The attacker estimates an approximation to the full record r as a function of $aux(r)$ and shared data $\mathbb{D}'$; thus, estimated record $r' := A(\mathbb{D}',aux(r))$. where, A is the adversary function the attacker uses to implement re-identification. A record is re-identified if the estimated and original record are in $\epsilon-$neighborhood of each other. The number of records re-identified correctly is a metric of success of the attack. Here, we assume that the auxiliary information is in the form of knowledge of certain quasi-identifiers and using the shared dataset, attacker wants to estimate the remaining quasi-identifiers. Also, we assume that the adversary function here is a multi-layer perceptron regressor. Accurate knowledge of the entire tuple of quasi-identifiers results in unique re-identification of individuals corresponding to the tuple of quasi-identifiers. We randomly sample 1% of records (325) as the re-identification targets of the attacker. The adversary strength is varied by varying the number of quasi-identifiers known to the attacker through the auxiliary information. The threshold is $\epsilon = 3$. The results of this attack in Figure 2 (right) show that the success rate is less than 10%, which can be further reduced by increasing the privacy weights.

Attribute disclosure Attack Model: In attribute disclosure attack, attacker uses auxiliary knowledge about certain attributes. We randomly sample a subset of the original data as the *compromised records.* We assume that attacker has knowledge about quasi-identifier attributes for these records and using the shared data the attacker wants to estimate the value of sensitive attribute. Here we consider *race.* For each record x, the attacker performs k-nearest neighbor classification to predict race of the user. We evaluate the effectiveness of attack through precision of the k-NN regression. For k = 5, Figure 2 (left) summarizes the results. We observe similar pattern for other values of $k \in \{1,2,..,10\}$. Two values of *race* are dominant in the data and so a random guess will have close to 50% accuracy; the generated data also shows similar behavior where an adversary can't perform better than a random guess. This depicts a strong performance of our approach against disclosure attack. In summary, the two attack models show low performance scores for the privacy-aware shared data and hence

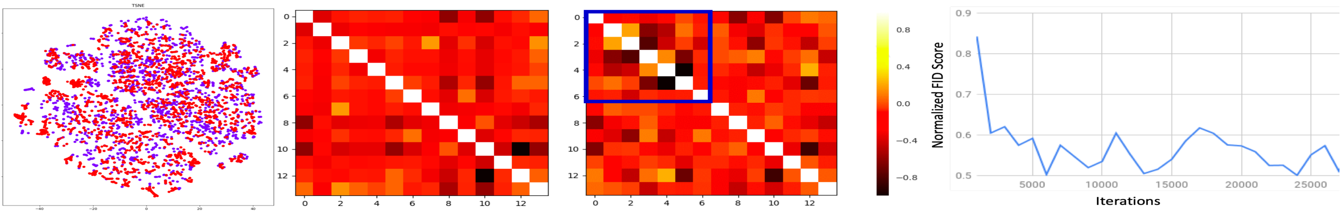

Figure 3: Quality of generated data: (L) t-SNE plot of original (purple) vs. generated (red) data. (C) Attribute correlation matrix of original vs. generated data. (R) Normalized FID score between generated and original data stabilizes after 5000 iterations.

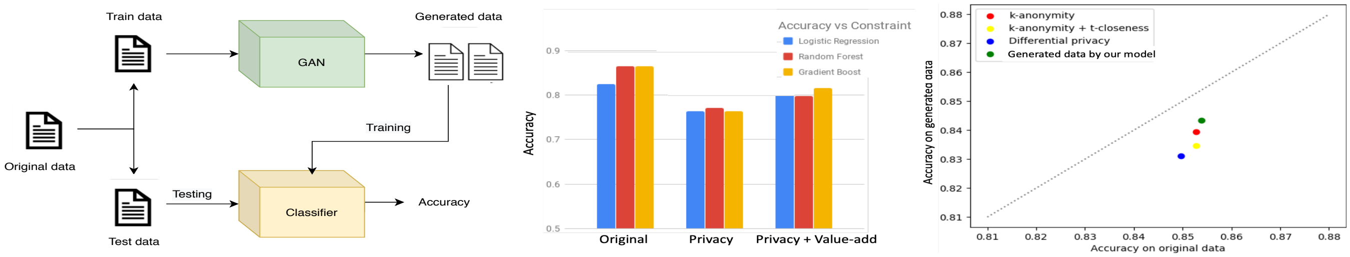

Figure 4: Evaluation of Value-addition: (L) Classification experiment setup. (C) Results on classification: We achieve accuracy close to original data which explains that generated data retains the value-addition for business. (R) Results on logistic regression; x is based on training on original data and testing on a subset of unseen original data, y is obtained by training on synthetic data and testing on the same unseen subset of original data.

ensures privacy. Next we show that generated data, although privacy protected, achieves necessary value addition for designated tasks.

Experimentation for Value-addition: We use a binary classification task. The experimental setup is depicted on the left hand side of Figure 4. We split the original data into train and test sets. We train the generative model using the train set and the classifier using the generated data. Performance is measured by AUC and accuracy on the test set. Results in Figure 4 (center) show that, for each model - Logistic Regression, Random Forest and Gradient Boosted - accuracy reduces from original data to that on data generated with *only* privacy regularizer. However, on data generated with *both* value-addition and privacy regularizers, the accuracy improves for each model. We evaluate value-addition as performance on the classification task with other approaches like k-anonymity, t-closeness and differential privacy, using the ARX de-identification tool [9]. We use $k = 5$ for anonymization, $t = 0.01$ for t-closeness and $\epsilon = 1.0, \delta = 10^-6$ for (ϵ, δ)- differential privacy. Figure 4 (right) shows the plot where x-axis is the classification metric obtained by training on original data and testing on a subset of unseen original data, and y-axis is the same metric obtained by training on generated data and testing on the same unseen subset of original data. An ideal anonymization / perturbation data generation scheme should result in points on the line $x = y$. Our approach performs better than the baselines.

4 CONCLUSION

Relevant to the accelerating data-sharing economy, we propose an approach that leverages generative model to generate representational data for sharing, which are not only privacy-protected but also tailored for value addition tasks of the recipient. Our method preserves the dependency structure among non-sensitive attributes unlike previous approaches which can lose associations in the data. Evaluations show that our dual regularizer-based workflow can improve the value added by the shared data for the recipient, compared to data generated with only privacy in mind. Yet we provide strong defense against advanced privacy attacks. Notably, the extent of privacy and value-addition are weights that can be tuned to influence generation of data, providing data provider with the flexibility in setting different degrees of trade-off between value-addition and privacy.

REFERENCES

[1] Rui Chen, Qian Xiao, Yu Zhang, and Jianliang Xu. 2015. Differentially Private High-Dimensional Data Publication via Sampling-Based Inference. In *KDD*. -, -.
[2] Simon Field. 2019. How to capitalise on the power of modern data sharing. https://www.information-age.com/capitalise-power-modern-data-sharing-123481528/.
[3] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *NIPS*. -, Long Beach, –.
[4] Jeffrey Mervis. 2019. Privacy concerns could derail Facebook data-sharing plan.
[5] Larry Myler. 2017. Data Sharing Can Be A Catalyst For B2B Innovation. https://www.forbes.com/sites/larrymyler/2017/09/11/data-sharing-can-be-a-catalyst-for-b2b-innovation/.
[6] Arvind Narayanan and Vitaly Shmatikov. 2006. How To Break Anonymity of the Netflix Prize Dataset. *CoRR* -, - (2006), –. http://arxiv.org/abs/cs/0610105
[7] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data Synthesis Based on Generative Adversarial Networks. In *VLDB*. -, -, –.
[8] Thierry Pellegrino. 2019. Protecting Data Privacy in a Data-Sharing World. https://www.cio.com/article/3393966/protecting-data-privacy-in-a-data-sharing-world.html.
[9] Fabian Prasser, Florian Kohlmayer, Gkoulalas-Divanis, Aris, and Grigorios Loukides. 2015. *Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool*. Springer, -, 111–148.
[10] Latanya Sweeney. 2002. K-anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* -, - (Oct. 2002), 557–570.
[11] UCI. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[12] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. In *NeurIPS*. -, Vancouver, –.
[13] Lei Xu and Kalyan Veeramachaneni. 2018. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264* -, - (2018), –.
[14] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private Data Release via Bayesian Networks. *ACM Trans. Database Syst.* -, - (Oct. 2017), –.