Abstract

Differentially-private mechanisms for text generation typically add carefully calibrated noise to input words and use the nearest neighbor to the noised input as the output word. When the noise is small in magnitude, these mechanisms are susceptible to reconstruction of the original sensitive text. This is because the nearest neighbor to the noised input is likely to be the original input. To mitigate this empirical privacy risk, we propose a novel class of differentially private mechanisms that parameterizes the nearest neighbor selection criterion in traditional mechanisms. Motivated by Vickrey auction, where only the second highest price is revealed and the highest price is kept private, we balance the choice between the first and the second nearest neighbors in the proposed class of mechanisms using a tuning parameter. This parameter is selected by empirically solving a constrained optimization problem for maximizing utility, while maintaining the desired privacy guarantees. We argue that this empirical measurement framework can be used to align different mechanisms along a common benchmark for their privacy-utility trade-off, particularly when different distance metrics are used to calibrate the amount of noise added. Our experiments on real text classification datasets show up to 50% improvement in utility compared to the existing state-of-the-art with the same empirical privacy guarantee.

1 Introduction

Over the past decade, privacy-preserving machine learning has emerged as a hot topic in a variety of real world speech and language applications. In natural language processing (NLP), ensuring data privacy in machine learning tasks is especially challenging because text data tends to be rich in sensitive and potentially identifiable information about the users that contributed to these datasets.

The literature is replete with approaches proposed for privacy-preserving text analysis, such as replacing sensitive information with general terms (Cumby and Ghani, 2011; Anandan et al., 2012; Sánchez and Batet, 2016), injecting additional words into original texts (Domingo-Ferrer et al., 2009; Pang et al., 2010; Sánchez et al., 2013), as well as k-anonymity and its variants (Sweeney, 2002; Machanavajjhala et al., 2007; Li et al., 2007). However, these methods are provably non-private and have been shown to be vulnerable to re-identification attacks (Korolova et al., 2009; Petit et al., 2015). To ensure a quantifiable privacy guarantee, differential privacy (DP) has become the de facto standard for privacy-preserving statistical analysis (Dwork et al., 2006; Dwork, 2008; Dwork et al., 2014), with applications to text analysis.

At a high level, a randomized algorithm is differentially private if the output distributions from any two neighboring databases are (near) indistinguishable. This indistinguishability is controlled by a privacy parameter, which, in the case of text analysis, is often scaled by the distance between neighboring datasets to capture the semantic similarity between different words (Feyisetan et al., 2019; Fernandes et al., 2019; Feyisetan et al., 2020; Xu et al., 2020). This calibration enables the mechanisms to enjoy metric-DP (Andrés et al., 2013; Chatzikokolakis et al., 2013), which was first introduced as a generalization of local DP (Kasiviswanathan et al., 2011) for protecting location privacy. Observe that a direct application of local DP mechanisms will be too restrictive because it requires that the probability ratio between the output distributions of any two words in the vocabulary be bounded by some fixed constant. Due to the high dimensional nature of textual tasks and very large vocabulary sizes (e.g. 2.2M words for GloVe common crawl (Pennington et al., 2014)), this can lead to adding a lot of noise for achieving the desired privacy guarantees, severely impacting the utility of the NLP task.

Comparing Metric-DP Mechanisms. In the context of text analysis, we are given a vocabulary...
set $\mathcal{W}$ and an embedding function $\phi : \mathcal{W} \rightarrow \mathbb{R}^p$, where $p$ is the dimensionality of the embedding model. For any $\epsilon > 0$, a mechanism $M : \mathcal{W} \rightarrow \mathcal{W}$ is said to be $\epsilon$ differentially private with respect to a given metric $d : \mathbb{R}^p \times \mathbb{R}^p \rightarrow [0, \infty)$ if for any $w, w', \hat{w} \in \mathcal{W}$, the following holds:

$$\Pr\{M(w) = \hat{w}\} \leq e^{\epsilon d(\phi(w), \phi(w'))} \Pr\{M(w') = \hat{w}\}. \tag{1}$$

The probabilistic guarantee in (1) ensures that the log probability ratio of observing any output $\hat{w}$ given two inputs $w$ and $w'$ is bounded by $\epsilon d(\phi(w), \phi(w'))$. This makes metric-DP less restrictive than the original notion of local DP (Kasiviswanathan, et al., 2013; Laud et al., 2020), as it provides meaningful privacy guarantees. This is because the pairwise distance in the word embedding space can be small floating numbers, which brings $\exp (30d(\phi(w), \phi(w')))$ to a reasonable scale. Thus, $\epsilon$ alone cannot fully characterize the privacy guarantee without the knowledge of the underlying metric space. More importantly, this indicates that the privacy guarantees from DP mechanisms with respect to different metrics are not directly comparable using only their $\epsilon$ values.

Our Contributions. A common feature in the existing metric-DP text generation mechanisms is to add a calibrate noise to the input word embedding and then output the nearest neighbor to the noisy embedding as the output. However, when the additive noise is small in magnitude, the input word is likely to remain unchanged, which may constitute an empirical privacy risk because it is trivial for the adversary to reconstruct the original word. To mitigate this issue, we present a novel class of metric-DP text generation mechanisms in this paper. Motivated by the Vickrey auction (Vickrey, 1961) scheme, also known as the second-price auction, we refer to this class of mechanisms as Vickrey mechanisms.

Just as in a Vickrey auction, where only the second highest price is revealed\(^1\) and the highest price is kept private, the proposed Vickrey mechanisms generalize the noisy nearest neighbor selection by including the second nearest neighbor in the selection pool using a tuning parameter. The inclusion of the second nearest neighbor greatly reduces the empirical reconstruction risk on the original word.

To select the tuning parameter above, we present a strategy based on optimizing the empirical privacy-utility tradeoff. The empirical privacy measurement is constructed in the context of analysis on de-identified text, which quantifies the risk on how well an adversary can reconstruct the original text based on the observed (possibly perturbed) text. The better the reconstruction, the lower the empirical privacy guarantee. This general framework allows comparing text generation mechanisms that use different distance metrics (see Section 3).

We emphasize that our empirical privacy metric does not supersede the metric-DP guarantee; instead, it provides a new dimension along which different metric-DP mechanisms can be aligned. We say that, within the class of metric-DP mechanisms, an optimal mechanism is the one that maximizes the empirical privacy guarantee while keeping the utility loss of the downstream task under some maximum tolerable budget. This definition for privacy-utility tradeoff, modeled as a constrained optimization problem, resembles the literature on protecting privacy for location data (Shokri et al., 2011, 2012; Clark et al., 2019). We extend the analysis for the broader class of metric-DP mechanisms. Additionally, in our experiments, we demonstrate that our proposed Vickrey mechanisms outperform existing mechanisms with respect to the empirical privacy-utility tradeoff on real text classification datasets.

Related Work. Metric-DP (Andrés et al., 2013; Chatzikokolakis et al., 2013; Laud et al., 2020), an extended notion of local DP (Kasiviswanathan et al., 2011), is a popular tool for privacy-preserving text analysis. A text generation mechanism that satisfies DP with respect to the hyperbolic distance metric was proposed in (Feyisetan et al., 2019). This mechanism requires specialized training of word embeddings in the high-dimensional hyperbolic space. For word embeddings in the Euclidean space, like GLOVe (Pennington et al., 2014) or FastText (Bojanowski et al., 2017), mechanisms like the Laplace mechanism ($L_2$ met-
ric) (Fernandes et al., 2019; Feyisetan et al., 2020) and the Mahalanobis mechanism (using a regularized Mahalanobis metric) (Xu et al., 2020) have been proposed. However, a structured comparison of these different mechanisms remains unclear.

Empirical privacy measurements. A variety of empirical techniques for privacy measurement have been proposed for many different applications. In the membership inference attack literature (Shokri et al., 2017; Yeom et al., 2018; Salem et al., 2018; Song and Shmatikov, 2019), an AUC based detectability metric is commonly used to quantify the information leakage from machine learning models about their training data. However, the model trained on a given dataset can only serve as a proxy to estimate its privacy guarantee. Moreover, the detectability metric can vary across different machine learning models and implementations of the inference attack based auditors.

Hypothesis testing based approaches have also been proposed to empirically estimate $\epsilon$ (Ding et al., 2018; Gilbert and McMillan, 2018; Liu and Oh, 2019). However, the assumptions in these methods constrain their general applicability. In a recent line of work on privacy-preserving text analysis (Feyisetan et al., 2020; Xu et al., 2020), privacy statistics defined as (i) probability of inputs not being redacted, and (ii) number of distinct outputs given a fixed input, have been used to characterize the empirical privacy of a text generation mechanisms. While those metrics are intuitive and descriptive, there is not a direct association that relates them to the privacy leakage. Within the class of metric-DP text generation mechanisms, the corresponding definition of empirical privacy-utility tradeoff is a constrained optimization to maximize the empirical privacy while keeping the utility loss under a preset budget. This constrained setup can find its precedent in the location data privacy literature (Shokri et al., 2011, 2012; Clark et al., 2019). We differ in their approach as we require the optimal mechanism to also satisfy metric-DP.

2 The Class of Vickrey Mechanisms

To motivate our construction of the Vickrey mechanisms, we begin by discussing the limitations of a general approach in the existing metric DP text generation mechanisms. We denote by $\mathcal{W} = \{ w_1, \ldots, w_n \}$ the vocabulary set containing $n$ distinct words, and by $\phi : \mathcal{W} \to \mathbb{R}^p$ a fixed embedding function that maps each word in the vocabulary set to a $p$-dimensional real vector (referred to as the embedding for the word).

A common first step is to sample an additive noise $Z$ from a density function $p(z) \propto \exp\{-d(z, 0)\}$, where $d$ is the distance metric used in the mechanism. For example, the Laplace mechanism uses $d(x, y) = ||x - y||_2$ (also known as Euclidean or $L_2$ distance), and the Mahalanobis mechanism uses $d(x, y) = \sqrt{(x - y)\Sigma^{-1}(x - y)}$ (also known as Mahalanobis distance), where $\Sigma$ is the sample covariance of the word embeddings.

Once the noise is sampled, it is then added to the input word embedding and the word with an embedding that is nearest to this noised embedding is chosen as the output:

$$w_{\text{output}} = \arg \min_{w \in \mathcal{W}} d(\phi(w_{\text{input}}) + Z, w).$$

A limitation of this noisy nearest neighbor selection is that when $|Z|$ is small (in particular, smaller than half the distance from the input word to its nearest neighbor), the first nearest neighbor to the noised embedding is the same as the original input word. The problem is exaggerated for rare words, which exist in the sparse regions of the embedding space and hence, do not get perturbed even for larger noise scales. This makes it easier for an adversary to reconstruct the original word, which may contain sensitive information (e.g., street names).

The proposed Vickrey mechanisms generalize the noisy nearest neighbor selection step by distributing the selection probability between the first and second nearest neighbor using a tuning parameter $t \in [0, 1]$ (see Algorithm 1). Intuitively, this generalization makes the reconstruction of the original input word harder (see Figures 1 and 2).

We capture our intuition for the claim above in Figure 1. For simplicity, the horizontal axis in both plots represents the one-dimensional embedding on a vocabulary containing only 5 words: (A, B, C, D, E). The vertical axis represents the output probability of each word through the mechanism. The plots represent the output probability in the mechanism for each of the 5 words, corresponding to the potential noised embedding values on the horizontal axis. The top plot represents the Laplace mechanism when only the first nearest neighbor
We formally prove that the Vickrey mechanism will select both the first and second nearest neighbors.

Figure 1: Word probability in the Laplace mechanism (top) and the Vickrey mechanism (bottom) at $0 < t < 1$ for each of the 5 words as a function of the noised one-dimensional embedding. The Vickrey mechanism in this example always has two candidate words as output.

Algorithm 1: The Vickrey Mechanism

Input: String $s = w_1 w_2 \ldots w_n$, metric $d$, privacy parameter $\epsilon$, tuning parameter $t \in [0, 1]$

for $w_i \in s$ do

1. Sample $Z$ with density $p(z) \propto \exp \{- \epsilon d(z, 0)\}$.
2. Obtain $\tilde{w}_i \leftarrow \phi(w_i) + Z$.
3. Let $\hat{w}_i \leftarrow \arg \min_{w \in W \setminus \{w_i\}} \|\tilde{w}_i - \phi(w)\|_2$, and
   $\tilde{w}_i \leftarrow \tilde{w}_i - (1 - t) \epsilon d(\tilde{w}_i - \phi(w_i)) / \|\tilde{w}_i - \phi(w_i)\|_2$.
4. $p(t, \hat{w}_i) = \frac{(1 - t) \|\tilde{w}_i - \phi(w_i)\|_2}{\|\tilde{w}_i - \phi(w_i)\|_2 + (1 - t) \|\tilde{w}_i - \phi(w_i)\|_2}$.

end for

return $\tilde{s} = \tilde{w}_1 \tilde{w}_2 \ldots \tilde{w}_n$.

Theorem 1. For any $t \in [0, 1]$, $\epsilon > 0$, metric $d$ and $w, w', \hat{w} \in W$, the Vickrey mechanism $M'_t$ from Algorithm 1 satisfies metric-DP:

$$\Pr\{M'_t(w) = \hat{w}\} / \Pr\{M'_t(w') = \hat{w}\} \leq \exp\left(\epsilon d(\phi(w), \phi(w'))\right).$$

Proof. Define $Q_{w_i}^{w_j} = \{v \in \mathbb{R}^p : \|v - \phi(w_i)\|_2 < \|v - \phi(w_j)\|_2 < \min_{w \in W \setminus \{w_i, w_j\}} \|v - \phi(w)\|_2\}$ to be the set that has $w_i$ and $w_j$ as the first and second nearest neighbors. Let $p_w(z)$ be the density function for the perturbed embedding conditional on the input $w$:

$$p_w(z) \propto \exp\left(-\epsilon d(z - \phi(w), 0)\right)$$

Since metric $d$ satisfies the triangle inequality,

$$d\{z - \phi(w'), 0\} - d\{z - \phi(w), 0\} \leq d\{\phi(w), \phi(w')\},$$

we obtain:

$$e^{-\epsilon d(z - \phi(w), 0)} \leq e^{\epsilon d(\phi(w), \phi(w'))} e^{-\epsilon d(z - \phi(w'), 0)},$$

which is equivalent to the inequality $p_w(z) \leq e^{\epsilon d(\phi(w), \phi(w'))} p_{w'}(z)$. For brevity, let

$$\alpha_{w'}^{w_j}(t, z) = \frac{1 - t}{t \|z - \phi(w_j)\|_2 + (1 - t) \|z - \phi(w_j)\|_2},$$

and $\rho(w, \hat{w}) = \Pr\{M'_t(w) = \hat{w}\}$. Since $\rho(w, \hat{w})$ is a sum of partial probabilities in the areas where $\hat{w}$ is either the first or the second nearest neighbor to the noised embedding, we have:

$$\rho(w, \hat{w}) = \sum_{j=1}^{|W|} \int_{Q_{w_j}^{w_i}} p_w(z) \alpha_{w_j}^{w_i}(t, z) dz + \sum_{i=1}^{|W|} \int_{Q_{w_i}^{w_j}} p_{w'}(z) \{1 - \alpha_{w_j}^{w_i}(t, z)\} dz \leq C(w, w') \left[ \sum_{j=1}^{|W|} \int_{Q_{w_j}^{w_i}} p_w(z) \{1 - \alpha_{w_j}^{w_i}(t, z)\} dz + \sum_{i=1}^{|W|} \int_{Q_{w_i}^{w_j}} p_{w'}(z) \{1 - \alpha_{w_j}^{w_i}(t, z)\} dz \right] = C(w, w') \Pr\{M'_t(w') = \hat{w}\},$$

where $C(w, w') = e^{\epsilon d(\phi(w), \phi(w'))}$, as desired. $\square$

For our experiments, we use the Euclidean distance for $d$ so that the Vickrey mechanism reduces to the Laplace mechanism when $t = 0$. In general, any distance function $d$ that satisfies the triangle inequality can be used to ensure the desired metric-DP guarantee (quantified by the parameter $\epsilon$).
3 Tuning Parameter Selection

We now discuss how to select the tuning parameter in Algorithm 1. We do this by optimizing an empirical formulation of the privacy-utility tradeoff. We discuss the details of this formulation next.

3.1 General Framework for Empirical Privacy Utility Tradeoff

Let \( M : W \rightarrow W \) denote some privacy-preserving text generation mechanism (that maps words to their noised versions). Define \( f_M(w'|w) \triangleq \Pr\{M(w) = w'\} \) to be the probability of observing \( w' \) as the output of the mechanism \( M \) from the input word \( w \). Note that this probability is conditioned on the knowledge of \( w \). We assume a prior probability measure \( \pi : W \rightarrow [0, 1] \), which represents the adversary’s domain knowledge about the NLP task and distribution of words in the dataset under consideration. Depending on the use case, the prior distribution \( \pi \) can be chosen as uniform, which means the user has no information on the word distribution in the context; or \( \pi \) can be chosen as the empirical word distribution in the corpus on which the user wishes to perform text generation.

Given this formulation, we define the expected utility loss for mechanism \( M \) as follows:

\[
L_M \triangleq \sum_{w,w'\in W} \pi(w)f_M(w'|w)d_L(w,w'),
\]

where \( d_L : W \times W \rightarrow [0, \infty) \) is a utility-specific distance metric. The utility loss can be bounded as \( L_M < C \) for some bound \( C > 0 \), depending on the maximum tolerance for the underlying task.

To model the empirical privacy loss, we assume an informative adversary \( A \) that uses the prior \( \pi \) and has full knowledge of the text generation mechanism \( M \) and the parameter \( \epsilon \) used (similar to (Shokri et al., 2011, 2012)). This adversary uses the posterior probability of each word given the observed perturbed output to make its inference:

\[
g_A(\hat{w}|w') \triangleq \frac{\pi(w')f_M(\hat{w}|w')}{\sum_{w\in W} \pi(w)f_M(w'|w')},
\]

Thus, from \( A \)'s perspective, the expected inference error with respect to \( M \) is given by:

\[
E_M = \sum_{w,w',\hat{w}} \pi(w)f_M(w'|w)g_A(\hat{w}|w)d_E(\hat{w}, w),
\]

where \( d_E : W \times W \rightarrow [0, \infty) \) is some privacy-specific distance metric. Our goal is, therefore, to find a mechanism within the class of metric-DP mechanisms \( M \) that maximizes the expected inference error \( E_M \) while keeping the utility loss \( L_M \) below \( C \):

\[
M_{optimal} = \arg \max_{M \in M} E_M, \ s.t. \ L_M < C. \tag{5}
\]

To compare different mechanisms, we will compare their expected inference error \( E_M \) under different tolerance thresholds on the expected utility loss \( L_M \). We favor mechanisms with high \( E_M \), while maintaining \( L_M < C \).

Note that \( d_L \) and \( d_E \) do not have to be the same distance metrics. For instance, \( d_L \) can depend on the downstream machine learning tasks, like the absolute difference in classification error, perplexity or even cross-entropy loss. From the privacy perspective, a natural choice is \( d_E(w, \hat{w}) = 1(\hat{w} \neq w) \), which means the adversary attempts to retrieve the original word from the redacted output and considers the inference attack successful if the inferred word is the exactly same as the input word. Based on applications, the adversary can also choose \( d_E \) to be the Euclidean distance such that the goal of the inference attack is to have the inferred word as close to the original word as possible.

3.2 Selecting the Tuning Parameter

We outline the main steps for optimizing the privacy parameter \( \epsilon \) as well as the tuning parameter \( t \) in Algorithm 2. This optimization is with respect to the empirical privacy-utility tradeoff as laid out in (5). We initialize with the privacy parameter \( \epsilon = \epsilon_0 \) at some small initial value \( \epsilon_0 \) and tuning parameter \( t = 0 \), so that the initial mechanism is essentially a metric-DP mechanism that implements the noisy first nearest neighbor selection. Next, we incrementally double the value of \( \epsilon \) until the expected
utility loss $L_{M^t} < C$ (recall that a smaller $\epsilon$ typically has larger utility loss\(^4\)). Once the maximum $\epsilon$ is obtained, we iterate over different values of $t$ between 0 and 1 (since a monotonicity assumption cannot be made here in general for the behavior of $E_M$). The final parameters $\epsilon_{\text{opt}}$ and $t_{\text{opt}}$ chosen provide the highest empirical privacy while keeping the utility loss within the specified budget. More importantly, Theorem 1 ensures that the selected mechanism enjoys at least as much metric DP as the initial mechanism, which implements only the nearest neighbor selection.

4 Experimental Results

Setup. We evaluate the performance of the proposed Vickrey mechanisms in terms of the empirical privacy-utility tradeoff on three datasets:

- The Product Reviews dataset consists of a list of 2,006 positive sentiment words and 4,783 negative sentiment words extracted from customer reviews (Hu and Liu, 2004). This is a word-level dataset and the metric $d_L$ in expected utility loss is $\mathbb{I}\{\text{sentiment}(w') \neq \text{sentiment}(w)\},$ i.e., the loss is incremented when a positive sentiment word is redacted into a negative sentiment word, or vice versa.

- The IMDb Movie Reviews dataset (Maas et al., 2011) has a total vocabulary size of 145,901, where a pre-specified set of 26,078 words are subject to redaction in the text generation mechanism (those are the words selected for adversarial model training in (Jia et al., 2019)). The utility task is the sentence-level binary sentiment classification, where the underlying model is a bidirectional LSTM using 90% of the data for training and 10% for testing.

- The Twitter dataset contains 7,613 tweets, with a vocabulary of 22,013 words\(^5\). Each tweet is associated with a label indicating whether the tweet describes a disaster event or not. The classification model is a bidirectional LSTM using 9:1 data split for training/testing.

For all three datasets, we consider both 300-dimensional GloVe embeddings (Pennington et al., 2014) and 300-d FastText embeddings (Bojanowski et al., 2017). The empirical privacy measurement uses the adversary’s expected inference error rate, i.e. $d_E(\hat{w}, w) = \mathbb{I}(\hat{w} \neq w)$. The utility-specific metric $d_L$ is chosen to be the misclassification error rate. The prior word distribution is chosen to be the empirical word distribution in the dataset, because we want to assume an informative adversary so as not to underestimate the privacy risk. In the Vickrey mechanism, the distance function is the Euclidean distance, so that $t = 0$ is equivalent to the Laplace mechanism (Feyisetan et al., 2020). We also compare our results with the Mahalanobis mechanism (Xu et al., 2020).

Results and Observations. In Figure 2(A) - 2(D) shows the empirical privacy-utility tradeoff on the Product Reviews between the Laplace mechanism, Mahalanobis mechanism, and the Vickrey mechanisms with tuning parameter at to 0.25, 0.5, 0.75, and 1. The vertical axis in all plots represents the adversary’s inference error in the mechanism. The error bars are computed over 100 runs. In the 2(A), the horizontal axis is the privacy budget $\epsilon$. When $\epsilon$ approaches 0, the inference error in all mechanisms approach 1, which is expected because magnitude of the additive noise is large. When $\epsilon$ increases, the inference error drops, but the drop in Laplace mechanism is much faster than the other mechanisms. It is worth noticing that the curves for Laplace mechanism and the Vickrey mechanisms are mostly parallel with each other: when $t$ increases from 0 to 0.75, a higher value of $t$ is better in terms of empirical privacy at the same $\epsilon$; but when $t$ increases to 1, the empirical privacy will not further increase since the randomness in noisy selection between the first and second nearest neighbor is replaced by the deterministic selection of the second nearest neighbor, which makes the adversary’s inference attack easier by finding the second nearest neighbor. However, Vickrey mechanism at $t = 1$ still dominates Laplace mechanism which only selects the noisy first nearest neighbor for redaction. The slope for the Mahalanobis mechanism is different from the rest, where intersects with Vickrey mechanisms with different $t$ at different $\epsilon$. At $\epsilon = 100$, the baseline Laplace mechanism has negligible inference error, which means the adversary can almost always make correct guesses, whereas in the other mechanisms the error is still substantial.

Figure 2(B) plots inference error vs. misclassifi-
Figure 2: (A): empirical privacy vs $\epsilon$ on Product Reviews using 300-d GLOVE. (B): empirical privacy vs utility loss on Product Reviews using 300-d GLOVE. (C): empirical privacy vs $\epsilon$ on Product Reviews using 300-d FASTTEXT. (D): empirical privacy vs utility loss on Product Reviews using 300-d FASTTEXT. (E) - (H) are for IMDb reviews, and (I) - (L) are for Twitter dataset.

fication error of words (positive words to negative words, or vice versa). When vertically slicing the plot, we see that for each utility loss budget greater than 0.1, a larger value of $t$ will result in a better privacy guarantee. When capped at a maximum $\epsilon = 100$, the curves with a higher $t$ value will have a higher minimum feasible misclassification error, which is around around 0.02 for $t = 0.25$ and Mahalanobis, about 0.04 for $t = 0.5$, about 0.06 for $t = 0.75$, and about 0.1 for $t = 1$. This is expected because as more weight is put on the second nearest neighbor, the utility loss becomes larger at large $\epsilon$ values (small noise), because it is more likely that the original word will get changed to its neighbors. But this loss is upper bounded by the nearest neighbor replacement, which tends to be small as is shown in the experiments in the paper. The plot suggests that if the user has a maximum utility loss budget of 0.06, they should go with the Vickrey mechanism at $t = 0.75$ because when slicing vertically at misclassification error of 0.06, the green curve for $t = 0.75$ attains a higher empirical privacy than the other mechanisms. However, when the utility loss budget is 0.02, the user should choose $t = 0.25$ because the green line is on top of the other curves (red and black) that can achieve the utility loss of 0.02 (the blue, cyan, and purple curves cannot achieve utility loss within 0.02 when
The difference in the result between GloVe and FASTTEXT, particularly in 2(E) vs. 2(G) and 2(I) vs. 2(K), is due to the difference in inter-word views and in Figure 2(I) - 2(L) on Twitter. In empirical privacy vs $\epsilon$ plots, the Laplace mechanism consistently has a lower value of empirical privacy measure than the Vickrey mechanism and the Mahalanobis mechanism. This gap in adversary’s inference error becomes wider as $\epsilon$ increases. In the privacy vs utility loss plots, the difference between mechanisms is more significant on Twitter than on IMDb reviews. The patterns are consistent across plots, which both show that the Vickrey mechanism can improve the privacy-utility tradeoff beyond the baseline mechanism.

The difference in the result between GloVe and FASTTEXT, particularly in 2(E) vs. 2(G) and 2(I) vs. 2(K), is due to the difference in inter-word distance distributions between the two embedding spaces (see Figure 1 in (Feyisetan et al., 2020)). In particular, the inter-word distances are generally smaller in FASTTEXT than in GloVe, so that for a fixed noise scale $\epsilon$, the inference error is expected to be larger in FASTTEXT than in GloVe.

5 Generalizing Vickrey Mechanism

5 Generalizing Vickrey Mechanism

Beyond the Second Nearest Neighbor

By a random selection of both the first and the second nearest neighbor to the noised embedding, we have shown that the Vickrey mechanism can empirically improve the privacy-utility tradeoff upon the existing Laplace and Mahalanobis mechanisms. A natural generalization is to extend the selection to $k \geq 2$ nearest neighbors (see Algorithm 3).

Algorithm 3 presents the outline of the generalized Vickrey mechanism that randomly chooses among the noisy $k$ nearest neighbors as output, where the selection probability is inversely associated with their distance to the noised embedding. Similar to Algorithm 2, the tuning parameters are selected as to optimize for the empirical privacy-utility tradeoff, but the selection process will be more challenging because the optimization space is unbounded. We defer the details of this optimization to future work. However, we formally state in Theorem 2 the metric-DP guarantee from the generalized Vickrey mechanism in Algorithm 2. Due to space constraints, we defer the details of the proof since it is similar to that for Theorem 1.

Theorem 2. For any $t = [t_1, \ldots, t_k] \in [0, \infty]^k$, $k \in \mathbb{Z}_+$, $\epsilon > 0$ and $w, w', \hat{w} \in W$, the generalized Vickrey mechanism $\mathcal{M}_t^\epsilon$ from Algorithm 3 satisfies $\epsilon$ metric-DP for any metric $d$.

In Figure 3, we compare 5 generalizations of the Vickrey mechanism that deterministically select the noisy $1^{st}, \ldots, 5^{th}$ neighbors as the output on the Product Reviews data using both 300-d GloVe and 300-d FASTTEXT. We can see that the improvement is most significant between the $1^{st}$ and $2^{nd}$ nearest neighbor. It also shows that there is benefit in introducing the $3^{rd}$ nearest neighbor into the selection pool, while no big difference is found beyond the $3^{rd}$ neighbor.
6 Discussion and Conclusion

In this paper, we present a measurement framework to quantify the empirical privacy-utility tradeoff for metric-DP text generation mechanisms, where the empirical privacy metric is the reconstruction risk of the original text based on the redacted text. We adopt a constrained optimization setup, where within the class of metric-DP mechanisms, we maximize the empirical privacy guarantee while keeping the machine learning utility loss under a pre-specified tolerance. A novel class of Vickrey mechanism is proposed, which not only enjoys metric-DP but also optimizes the privacy-utility tradeoff within the constraint. We apply our methodology to the three text classification datasets and demonstrate how to empirically compare the privacy-utility tradeoff as well as how to choose the optimal parameter setting according to the constrained optimization. Our results show superior performance when compared to existing mechanisms.

Our analysis in this paper leaves ample room for further investigation. An ongoing work we are exploring is the inclusion of contextual information into the probability calibration between the two nearest neighbors. We leave it as an interesting open problem to explore how the choice of $k^{th}$ neighbor impacts the tradeoff in this scenario, since contextual signals will likely restrict the set of candidate words we can choose from.

References


Josep Domingo-Ferrer, Agusti Solanas, and Jordi Castellà-Roca. 2009. h (k)-private information retrieval from privacy-uncooperative queryable databases. Online Information Review.


