A Differentially Private Text Perturbation Method Using a Regularized Mahalanobis Metric

Zekun Xu
Amazon
Seattle, WA USA
zeku@amazon.com

Abhinav Aggarwal
Amazon
Seattle, WA USA
aggabhin@amazon.com

Oluwaseyi Feyisetan
Amazon
Seattle, WA USA
sey@amazon.com

Nathanael Teissier
Amazon
Arlington, VA USA
natteis@amazon.com

Abstract

Balancing the privacy-utility tradeoff is a crucial requirement of many practical machine learning systems that deal with sensitive customer data. A popular approach for privacy-preserving text analysis is noise injection, in which text data is first mapped into a continuous embedding space, perturbed by sampling a spherical noise from an appropriate distribution, and then projected back to the discrete vocabulary space. While this allows the perturbation to admit the required metric differential privacy, often the utility of downstream tasks modeled on this perturbed data is low because the spherical noise does not account for the variability in the density around different words in the embedding space. In particular, words in a sparse region are likely unchanged even when the noise scale is large.

In this paper, we propose a text perturbation mechanism based on a carefully designed regularized variant of the Mahalanobis metric to overcome this problem. For any given noise scale, this metric adds an elliptical noise to account for the covariance structure in the embedding space. This heterogeneity in the noise scale along different directions helps ensure that the words in the sparse region have sufficient likelihood of replacement without sacrificing the overall utility. We provide a text-perturbation algorithm based on this metric and formally prove its privacy guarantees. Additionally, we empirically show that our mechanism improves the privacy statistics to achieve the same level of utility as compared to the state-of-the-art Laplace mechanism.

1 Introduction

Machine learning has been successfully utilized in a wide variety of real world applications including information retrieval, computer graphics, speech recognition, and text mining. Technology companies like Amazon, Google, and Microsoft already provide MLaaS (Machine Learning as a Service), where customers can input their datasets for model training and receive black-box prediction results as output. However, those datasets may contain personal and potentially sensitive information, which can be exploited to identify the individuals in the datasets, even if it has been anonymized (Sweeney, 1997; Narayanan and Shmatikov, 2008). Removing personally identifiable information is often inadequate, since having access to the summary statistics on the dataset has been shown to be sufficient to infer individual's membership in the dataset with high probability (Homer N, 2008; Sankararaman S., 2009; Dwork et al., 2015). Moreover, machine learning models themselves can reveal information on the training data. In particular, sophisticated deep neural networks for natural language processing tasks like next word prediction or neural machine translation, often tend to memorize their training data, which makes them vulnerable to leaking information about their training data (Shokri et al., 2017; Salem et al., 2018).

To provide a quantifiable privacy guarantee against such information leakage, Differential Privacy (DP) has been adopted as a standard framework for privacy-preserving analysis in statistical databases (Dwork et al., 2006; Dwork, 2008; Dwork et al., 2014). Intuitively, a randomized algorithm is differentially private if the output distributions from two neighboring databases are indistinguishable. However, a direct application of DP to text analysis can be too restrictive because it requires a lower bound on the probability of any word to be replaced by any other word in the vocabulary.

Metric differential privacy arises as a generalization of local differential privacy (Kasiviswanathan et al., 2011), which originated in protecting location privacy such that locations near the user's location are assigned with higher probability while those far away are given negligible probability

Proceedings of the Second Workshop on Privacy in Natural Language Processing, pages 7–17

https://doi.org/10.18653/v1/P17

(Andrés et al., 2013; Chatzikokolakis et al., 2013). In the context of privacy-preserving text analysis, metric differential privacy implies that the indistinguishability of the output distributions of any two words in the vocabulary is scaled by their distance, where the distance metrics used in the literature include Hamming distance (reduced to DP), Manhattan distance (Chatzikokolakis et al., 2015), Euclidean distance (Chatzikokolakis et al., 2013; Fernandes et al., 2019; Feyisetan et al., 2020), Chebyshev distance (Wagner and Eckhoff, 2018), hyperbolic distance (Feyisetan et al., 2019).

In this paper, we propose a novel privacy-preserving text perturbation method by adding an elliptical noise to word embeddings in the Euclidean space, where the scale of the noise is calibrated by the regularized Mahalanobis norm (formally defined in Section 3). We compare our method to the existing multivariate Laplace mechanism for privacy-preserving text analysis in the Euclidean space (Fernandes et al., 2019; Feyisetan et al., 2020). In both papers, text perturbation is implemented by adding a spherical noise sampled from multivariate Laplace distribution to the original word embedding. However, the spherical noise does not account for the structure in the embedding space. In particular, words in a sparse region are likely unchanged even when the scale of noise is large. This can potentially result in severe privacy breach when sensitive words do not get perturbed. To increase the substitution probability of words in sparse regions, the scale of noise has to be large in the multivariate Laplace mechanism, which will hurt the downstream machine learning utility.

We address this problem by adding an elliptical noise to word embeddings according to the covariance structure in the embedding space. The intuition is that given a fixed scale of noise, we want to stretch the noise equidistant contour in the direction so that the substitution probability of words in the sparse region is increased on average. Intuitively, this direction is the one that explains the largest variability in the word embedding vectors in the vocabulary. We prove the theoretical metric differential privacy guarantee of the proposed method. Furthermore, we use empirical analysis to show that the proposed method significantly improves the privacy statistics while achieves the same level of utility as compared to the multivariate Laplace mechanism. Our main contributions are as follows:

- We develop a novel Mahalanobis mechanism for differentially private text perturbation, which calibrates the elliptical noise by accounting for the covariance structure in the word embedding space.
- A theoretical metric differential privacy proof is provided for the proposed method.
- We compare the privacy statistics and utility results between our method and the multivariate Laplace Mechanism through experiments, which demonstrates that our method has significantly better privacy statistics while preserving the same level of utility.

2 Related Works

Privacy-preserving text analysis is a well-studied problem in the literature (Hill et al., 2016). One of the common approaches is to identify sensitive terms (like personally identifiable information) in a document and replace them with some more general terms (Cumby and Ghani, 2011; Anandan et al., 2012; Sánchez and Batet, 2016). Another line of research achieves text redaction by injecting additional words into the original text without detecting sensitive entities (Domingo-Ferrer et al., 2009; Pang et al., 2010; SáNchez et al., 2013). However, those methods are shown to be vulnerable to reidentification attacks (Petit et al., 2015).

In order to provide a quantifiable theoretical privacy guarantee, the differential privacy framework (Dwork, 2008) has been used for privacy-preserving text analysis. In the DPText model (Beigi et al., 2019), an element-wise univariate Laplace noise is added to the pre-trained auto-encoders to provide privacy for text representations. Another approach for privacy-preserving text perturbation is in the metric differential privacy framework (Andrés et al., 2013; Chatzikokolakis et al., 2013), an extended notion of local differential privacy (Kasiviswanathan et al., 2011), which adds noise to the pre-trained word embeddings. Metric differential privacy requires that the indistinguishability of the output distributions of any two words in the vocabulary be scaled by their distance, which reduces to differential privacy when Hamming distance is used (Chatzikokolakis et al., 2013). A hyperbolic distance metric (Feyisetan et al., 2019) was proposed to provide privacy by perturbing vector representations of words, but it requires specialized training of word embeddings in the high-dimensional hyperbolic space. For the word em-

beddings in the Euclidean space (Fernandes et al., 2019; Feyisetan et al., 2020), text perturbation is implemented by sampling independent spherical noise from multivariate Laplace distributions. The former work (Fernandes et al., 2019) subsequently used an Earth mover's metric to derive a Bag-of-Words representation on the text, whereas the latter (Feyisetan et al., 2020) directly worked on the word-level embeddings. Since we work with word embeddings in the Euclidean space, we compare our method to the multivariate Laplace mechanism for text perturbation in those two papers (Fernandes et al., 2019; Feyisetan et al., 2020).

Mahalanobis distance has been used as a sensitivity metric for differential privacy in functional data (Hall et al., 2012, 2013) and differentially private outlier analysis (Okada et al., 2015). Outside the realm of text analysis, Mahalanobis distance is a common tool in cluster analysis, pattern recognition, and anomaly detection (De Maesschalck et al., 2000; Xiang et al., 2008; Warren et al., 2011; Zhao et al., 2015; Zhang et al., 2015).

3 Methodology

We begin by formally defining Euclidean norm and the regularized Mahalanobis norm.

Definition 1 (Euclidean Norm)**.** *For any vector $x \in \mathbb{R}^m$, its Euclidean norm is: $\|x\|_2 = \sqrt{x^\intercal x}$.*

Definition 2 (Mahalanobis Norm)**.** *For any vector $x \in \mathbb{R}^m$, and a positive definite matrix Σ, its Mahalanobis norm is:*

$$\|x\|_M = \sqrt{x^\intercal \Sigma^{-1} x}.$$

Definition 3 (Regularized Mahalanobis Norm)**.** *For any vector $x \in \mathbb{R}^m$, $\lambda \in [0, 1]$, and a positive definite matrix Σ, its regularized Mahalanobis norm is:*

$$\|x\|_{RM} = \sqrt{x^\intercal \{\lambda\Sigma + (1-\lambda) I_m\}^{-1} x}.$$

From the definitions above, λ can be considered as a tuning parameter. When $\lambda = 0$, the regularized Mahalanobis norm reduces to the Euclidean norm; when $\lambda = 1$, the regularized norm reduces to the Mahalanobis norm (Mahalanobis, 1936)

Note that for any $\eta > 0$, the trajectory of $\{y \in \mathbb{R}^m : \|y - x\|_2 = \eta\}$ is spherical, whereas the trajectory of $\{y \in \mathbb{R}^m : \|y - x\|_{\mathsf{RM}} = \eta\}$ is elliptical unless $\lambda = 0$. We will exploit this key difference in the geometry of equidistant contour between the Euclidean norm and the regularized Mahalanobis norm to motivate our text perturbation method. The proposed regularized Mahalanobis norm is a type of shrinkage estimator (Daniels and Kass, 2001; Schäfer and Strimmer, 2005; Couillet and McKay, 2014), which is commonly used to estimate the covariance matrix of high-dimensional vectors so as to ensure the stability of the estimator. The matrix Σ in the regularized Mahalanobis norm controls the direction to which the equidistant contour in the noise distribution is stretched, while the parameter λ controls the degree of the stretch.

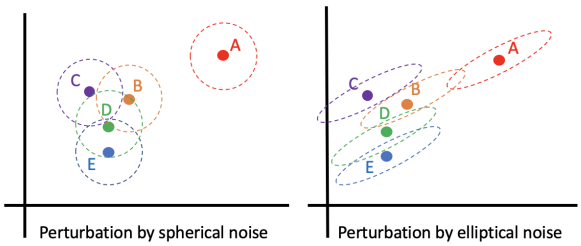

Figure 1: **Left:** dotted trajectories represent perturbation by spherical noise in multivariate Laplace mechanism. **Right:** dotted trajectories represent perturbation by elliptical noise in proposed Mahalanobis mechanism. The scale of noise is the same in both plots.

Definition 4 (Metric Differential Privacy.)**.** *For any $\epsilon > 0$, a randomized algorithm $M : \mathcal{X} \to \mathcal{Y}$ satisfies ϵ $d_{\mathcal{X}}-$privacy if for any $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$, the following holds:*

$$\frac{\Pr\{M(x) = y\}}{\Pr\{M(x') = y\}} \leq \exp\{\epsilon d(x, x')\}.$$

Metric differential privacy ($d_{\mathcal{X}}-$privacy) originated in privacy-preserving geolocation studies (Andrés et al., 2013; Chatzikokolakis et al., 2013), where the metric d is Euclidean distance. It has been extended to quantifying privacy guarantee in text analysis, which states that for any two words w, w' in the vocabulary $\mathcal{W}$, the likelihood ratio of observing any $\hat{w} \in \mathcal{W}$ is bounded by $\epsilon d(w, w')$, where $d(w, w') = \|\phi(w) - \phi(w')\|_2$ for an embedding function $\phi : \mathcal{W} \to \mathbb{R}^m$.

The multivariate Laplace mechanism (Fernandes et al., 2019; Feyisetan et al., 2020) perturbs the word embedding $\phi(w)$ by adding a random spherical noise Z sampled from density $f_Z(z) \propto e^{\epsilon\|z\|_2}$, and then find the nearest neighbor in the embedding space as the output of the mechanism.

The left panel in Figure 1 illustrates the text perturbation by spherical noise in the multivariate Laplace mechanism. Here A is in the sparse region in the two-dimensional embedding space.

Given the privacy budget ϵ, A has a small probability of being substituted by other words because its expected perturbation (dotted trajectory) still has itself as the nearest neighbor. The right panel in Figure 1 shows text perturbation in the proposed Mahalanobis mechanism, where the word embeddings are redacted by an elliptical noise at the same privacy budget ϵ. The matrix Σ is taken to be the sample covariance matrix of the word embeddings scaled by the mean sample variance, so that the noise contour is stretched toward the direction that explains the largest variability in the embedding space. The purpose of the scaling step is to ensure that the scale of the elliptical noise is the same as the scale of the spherical noise. By transforming the spherical noise contour into an elliptical contour, we increase the substitution probability of A in the sparse region, thus improving the privacy guarantee. Meanwhile, since the scale of noise does not change, the utility is preserved at the same level. This is an illustrative example demonstrating the intuition and motivation of the proposed Mahalanobis mechanism.

Our proposed Mahalanobis mechanism for text perturbation shares the same general structure as the multivariate Laplace mechanism. The key difference is that the spherical noise $f_Z(z) \propto \exp(\epsilon\|z\|_2)$ in the multivariate Laplace mechanism is replaced by the elliptical noise sampled from density $f_Z(z) \propto \exp(-\epsilon\|z\|_{\mathsf{RM}})$, which can be efficiently performed via Algorithm 1.

An overview for the proposed Mahalanobis mechanism is presented in Algorithm 2. When $\lambda = 0$, the proposed Mahalanobis mechanism reduces to the multivariate Laplace mechanism. A heuristic method for choosing the tuning parameter λ is to find the value of λ that maximizes the improvement in the privacy guarantee while maintaining the same level of utility. This can be done through empirical privacy and utility experiments as described in Section 5. The input Σ in Algorithm 2 is computed by scaling the sample covariance matrix of the word embeddings by the mean sample variance so as to guarantee that the trace of Σ equals the trace of I_m. Since Σ is a scaled counterpart of the sample covariance matrix, it will stretch the elliptical noise toward the direction with the largest variability in the word embedding space, which maximizes the overall expected probability of words being substituted. We remark that in order to maximize the substitution probability for each individual word, a personalized covariance matrix Σ_w can be computed in the neighborhood of each word. This is beyond the scope of this paper and we leave it as future work.

4 Theoretical Properties

Lemma 1. *The random variable Z returned from Algorithm 1 has a probability density function of the form $f_Z(z) \propto \exp(-\epsilon\|z\|_{\mathsf{RM}})$.*

Proof. Define $U = YX$. Note that conditional on Y, U follows a uniform distribution on a sphere with radius y in the $m-$dimensional space, which implies $f_{U|Y}(u|y) \propto 1/y^{m-1}$ when $\sum_{i=1}^m u_i^2 = y^2$ and 0 otherwise. Therefore,

$$\begin{aligned} f_U(u) &= \int f_{U|Y}(u|y) f_Y(y) \delta(y = \|u\|_2) dy \\ &\propto \int \frac{1}{y^{m-1}} \frac{\epsilon^m}{\Gamma(m)} y^{m-1} e^{-\epsilon y} \delta(y = \|u\|_2) dy \\ &\propto e^{-\epsilon\sqrt{u^\intercal u}}, \end{aligned}$$

where $\delta(\cdot)$ is the Dirac delta function. Since $Z = \{\lambda\Sigma + (1-\lambda)I_m\}^{1/2}U$, which is well-defined because $\lambda \in [0,1]$ and Σ is positive definite, it follows that:

$$f_Z(z) \propto \exp\left(-\epsilon\sqrt{z^\intercal\{\lambda\Sigma + (1-\lambda)I_m\}^{-1}z}\right),$$

The result in the lemma follows by definition. □

Theorem 1. *For any given $\epsilon > 0$ and $\lambda \in [0,1]$, the Mahalanobis mechanism from Algorithm 2 satisfies ϵ d_χ-privacy with respect to the Regularized Mahalanobis Norm.*

Proof. It suffices to show that for any strings $s = [w_1 \dots w_n]$, $s' = [w_1' \dots w_n']$, $\hat{s} = [\hat{w}_1 \dots \hat{w}_n]$, $\epsilon > 0$, $\lambda \in [0,1]$, and positive definite matrix Σ,

$$\frac{\Pr\{M(s) = \hat{s}\}}{\Pr\{M(s') = \hat{s}\}} \leq e^{\epsilon \sum_{i=1}^n \|\phi(w_i) - \phi(w_i')\|_{\mathsf{RM}}},$$

where $M : \mathcal{W}^n \to \mathcal{W}^n$ is the Mahalanobis mechanism and $\phi : \mathcal{W} \to \mathbb{R}^m$ is the embedding function.

We begin by showing that for any $w, w', \hat{w} \in \mathcal{W}$, it holds that the probability $\Pr\{M(w) = \hat{w}\}$ is at most $e^{\epsilon\|\phi(w) - \phi(w')\|_{\mathsf{RM}}}$ times the probability $\Pr\{M(w') = \hat{w}\}$. We define $C_{\hat{w}} = \{v \in \mathbb{R}^m : \|v - \phi(\hat{w})\|_2 < \min_{w \in \mathcal{W}\setminus\hat{w}} \|v - \phi(w)\|_2\}$ be the set of vectors v that are closer to $\hat{w}$ than any other

Algorithm 1: Sampling from $f_Z(z) \propto \exp(-\epsilon\|z\|_{\mathsf{RM}})$

Input: Dimension m, a positive definite matrix Σ, tuning parameter $\lambda \in [0, 1]$
Sample an m-dimensional random vector N from a multivariate normal distribution with mean zero and identity covariance matrix.
Normalize $X = N/\|N\|_2$.
Sample Y from a Gamma distribution with shape parameter m and scale parameter $1/\epsilon$.
Return $Z = Y \cdot \{\lambda\Sigma + (1-\lambda)I_m\}^{1/2}X$.

Algorithm 2: The Mahalanobis Mechanism

Input: String $s = w_1 w_2 \dots w_n$, privacy parameter $\epsilon > 0$, scaled sample covariance matrix Σ, tuning parameter $\lambda \in [0, 1]$
for $i \in \{1, \dots, n\}$ **do**
  Sample Z from $f_Z(z) \propto \exp(-\epsilon\|z\|_{\mathsf{RM}})$ using Algorithm 1.
  Obtain the perturbed embedding $\hat{\phi}_i = \phi(w_i) + Z$.
  Replace w_i with $\hat{w}_i = \arg\min_{w\in\mathcal{W}} \|\phi(w) - \hat{\phi}_i\|_2$.
return $\tilde{s} = \hat{w}_1\hat{w}_2 \dots \hat{w}_n$.

word in the embedding space. Let Z be sampled from $f_Z(z) \propto \exp(-\epsilon\|z\|_{\mathsf{RM}})$ by Algorithm 1,

$$\Pr\{M(w) = \hat{w}\} = \Pr\{\phi(w) + Z \in C_{\hat{w}}\}$$
$$= \int_{C_{\hat{w}}} f_Z\left(v - \phi(w)\right) dv = \int_{C_{\hat{w}}} e^{-\epsilon\|v-\phi(w)\|_{\mathsf{RM}}} dv,$$

where the last step follows from Lemma 1. Since Σ is positive definite, it admits a spectral decomposition $\Sigma = Q\Lambda Q^\intercal$, where $Q^\intercal = Q^{-1}$, and Λ is a diagonal matrix with positive entries $\xi_1, \dots, \xi_m$. Then we can rewrite $\{\lambda\Sigma + (1-\lambda)I_m\}^{-1} = Q\Omega Q^\intercal$, where $\Omega^{-1} = \lambda\Lambda + (1-\lambda)I_m$. Define $\tilde{v} = \Omega^{1/2}Q^\intercal v$ and $\tilde{\phi}(w) = \Omega^{1/2}Q^\intercal\phi(w)$. By the triangle inequality, the following hold:

$$\begin{aligned}
&e^{-\epsilon\sqrt{\{v-\phi(w)\}^\intercal\{\lambda\Sigma+(1-\lambda)I_d\}^{-1}\{v-\phi(w)\}}} \\
=&e^{-\epsilon\sqrt{\{v-\phi(w)\}^\intercal Q\Omega^{1/2}\Omega^{1/2}Q^\intercal\{v-\phi(w)\}}} \\
=&e^{-\epsilon\|\tilde{v}-\tilde{\phi}(w)\|_2} \\
=&\frac{e^{-\epsilon\|\tilde{v}-\tilde{\phi}(w')\|_2}}{e^{-\epsilon\|\tilde{v}-\tilde{\phi}(w')\|_2}} e^{-\epsilon\|\tilde{v}-\tilde{\phi}(w)\|_2} \\
\le&e^{-\epsilon\|\tilde{v}-\tilde{\phi}(w')\|_2} e^{\epsilon\|\tilde{\phi}(w)-\tilde{\phi}(w')\|_2} \\
=&e^{-\epsilon\|v-\phi(w')\|_{\mathsf{RM}}} e^{\epsilon\|\phi(w)-\phi(w')\|_{\mathsf{RM}}}.
\end{aligned}$$

The probability ratio is computed by:

$$\begin{aligned}
\frac{\Pr\{M(w) = \hat{w}\}}{\Pr\{M(w') = \hat{w}\}} &= \frac{\int_{C_{\hat{w}}} e^{-\epsilon\|v-\phi(w)\|_{\mathsf{RM}}} dv}{\int_{C_{\hat{w}}} e^{-\epsilon\|v-\phi(w')\|_{\mathsf{RM}}} dv} \\
&\le e^{\epsilon\|\phi(w)-\phi(w')\|_{\mathsf{RM}}}.
\end{aligned}$$

Finally, since each word in the string is processed independently,

$$\begin{aligned}
\frac{\Pr(M(s) = \hat{s})}{\Pr(M(s') = \hat{s})} &= \prod_{i=1}^{n}\left(\frac{\Pr(M(w_i) = \hat{w}_i)}{\Pr(M(w_i') = \hat{w}_i)}\right) \\
&\le \prod_{i=1}^{n} e^{\epsilon\|\phi(w_i)-\phi(w_i')\|_{\mathsf{RM}}} \\
&\le e^{\epsilon\sum_{i=1}^{n}\|\phi(w_i)-\phi(w_i')\|_{\mathsf{RM}}}. \square
\end{aligned}$$

Next, we relate the proved theoretical guarantee of the Mahalanobis mechanism to that of the multivariate Laplace mechanism (Fernandes et al., 2019; Feyisetan et al., 2020), which enjoys metrical differential privacy guarantee with respect to the Euclidean metric. The following lemma will help establish our result.

Lemma 2. *Let $v \in \mathbb{R}^2$ and $m = trace(\Sigma)$. Let $c > 0$ be a lower bound on the smallest eigenvalue of Σ. Then, the following bounds hold:*

$$\frac{||v||_2}{\sqrt{\lambda m + 1 - \lambda}} \le \|v\|_{\mathsf{RM}} \le \frac{||v||_2}{\sqrt{\lambda c + 1 - \lambda}}.$$

Proof. Since Σ is positive definite, it admits a spectral decomposition $\Sigma = Q\Lambda Q^\intercal$, where $Q^\intercal = Q^{-1}$, and Λ is a diagonal matrix with eigenvalues $\xi_1, \dots, \xi_m$. Since by assumption of the minimum eigenvalue greater than $c > 0$, and that $trace(\Sigma) = \sum_{i=1}^{m}\xi_i = m$, we have $\xi_i \in (c, m)$ for $i = 1, \dots, m$. Then, the eigenvalues for $\{\lambda\Sigma + (1-\lambda)I_m\}^{-1}$ are $\frac{1}{\lambda\xi_1+1-\lambda}, \dots, \frac{1}{\lambda\xi_m+1-\lambda}$,

which are between $\frac{\epsilon}{\lambda m+1-\lambda}$ and $\frac{\epsilon}{\lambda c+1-\lambda}$. Then for any vector $v \in \mathbb{R}^m$,

$$\begin{aligned}\|v\|_{\mathsf{RM}}^2 =& v^\intercal\{\lambda\Sigma + (1-\lambda)I_m\}^{-1}v \\ =& \sum_{i=1}^{m} \frac{1}{\lambda\xi_i + 1 - \lambda}(q_i^\intercal v)^2 \\ \leq& \frac{1}{\lambda c + 1 - \lambda}\sum_{i=1}^{m}(q_i^\intercal v)^2 = \frac{\|v\|_2^2}{\lambda c + 1 - \lambda},\end{aligned}$$

where $\sum_{i=1}^m (q_i^\intercal v)^2 = \|Q^\intercal v\|_2^2 = \|v\|_2^2$. Similarly, we can show that $\|v\|_{\mathsf{RM}}^2 \geq \frac{1}{\lambda m+1-\lambda}\|v\|_2^2$, so the result follows immediately. □

The lemma below then follows.

Lemma 3. *Assume $trace(\Sigma) = m$ and the minimum eigenvalue of Σ is greater than c for some constant $c > 0$, then for any $w, w' \in \mathcal{W}$, then*

$$\begin{aligned}&\exp\left(\epsilon\|\phi(w) - \phi(w')\|_{\mathsf{RM}}\right) \\ \leq& \exp\left(\frac{\epsilon}{\sqrt{\lambda c + 1 - \lambda}}\|\phi(w) - \phi(w')\|_2\right), \quad \textit{and} \\ &\exp\left(\epsilon\|\phi(w) - \phi(w')\|_{\mathsf{RM}}\right) \\ \geq& \exp\left(\frac{\epsilon}{\sqrt{\lambda m + 1 - \lambda}}\|\phi(w) - \phi(w')\|_2\right).\end{aligned}$$

The fact that the probability ratio as a function of $\|\cdot\|_{\mathsf{RM}}$ can be sandwiched by lower and upper bounds as functions of $\|\cdot\|_2$ shows the noise scale ϵ is comparable between the Mahalanobis mechanism and multivariate Laplace mechanism.

5 Experiments

We empirically compare the proposed Mahalanobis mechanism and the existing multivariate Laplace mechanism in both privacy experiments and utility experiments on the following two datasets (more details in Appendix A):

- **Twitter Dataset**. This is a publicly available Kaggle competition dataset (https://www.kaggle.com/c/nlp-getting-started), which contains 7,613 tweets, each with a label indicating whether the tweet describes a disaster event (43% disaster).
- **SMSSpam Dataset**. This is a publicly available dataset from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection), which contains 5,574 (13% spam) SMS labeled messages collected for mobile phone spam research (Almeida et al., 2011).

5.1 Privacy experiments

In the privacy experiments, we compare the Mahalanobis mechanism with the multivariate Laplace mechanism on the following two privacy statistics:

1. $N_w = \Pr\{M(w) = w\}$, which is the probability of the word not getting redacted in the mechanism. This is approximated by counting the number of times an input word w does not get replaced by other words after running the mechanism for 100 times.
2. $S_w = |\{w' \in \mathcal{W} : \Pr\{M(w) = w'\} \geq \eta|$, which is the number of distinct words that have a probability greater than η of being the output of $M(w)$. This is approximated by counting the number of distinct substitutions for an input word w after running the mechanism for 100 times.

We note that N_w and S_w have been previously used in privacy-preserving text analysis literature to qualitatively characterize the privacy guarantee (Feyisetan et al., 2019, 2020). We make the following connection between the privacy statistics (N_w and S_w) with the DP privacy budget ϵ: a smaller ϵ corresponds to a stronger privacy guarantee by adding a larger scale ($1/\epsilon$) of noise in the mechanism, which leads to fewer unperturbed words (lower N_w) and more diverse outputs for each word (higher S_w). For any fixed noise scale of $1/\epsilon$, the mechanism with a better privacy guarantee will have a lower value of N_w and higher value of S_w.

In Figure 2 and 3, we summarize how the 5^{th}, 50^{th}, and 95^{th} percentiles of N_w and S_w change in different configurations of the mechanisms using the 300-d FastText embedding (Bojanowski et al., 2017). The vocabulary set includes 28,596 words in the vocabulary union from the two real datasets. For all 5^{th}, 50^{th}, and 95^{th} percentiles, the Mahalanobis mechanism has a lower value of N_w and a higher value of S_w as compared to the multivariate Laplace mechanism, which indicates an improvement in privacy statistics.

Table 1 and 2 compare the mean and standard deviation of N_w and S_w across different settings. The mean N_w converges to 0 as ϵ decreases and converges to 100 as ϵ increases. An opposite trend is observed for S_w, which is as expected. In the middle range of privacy budget ($\epsilon = 5, 10, 20$), the proposed Mahalanobis mechanism has significantly lower values of N_w and higher values of S_w, where

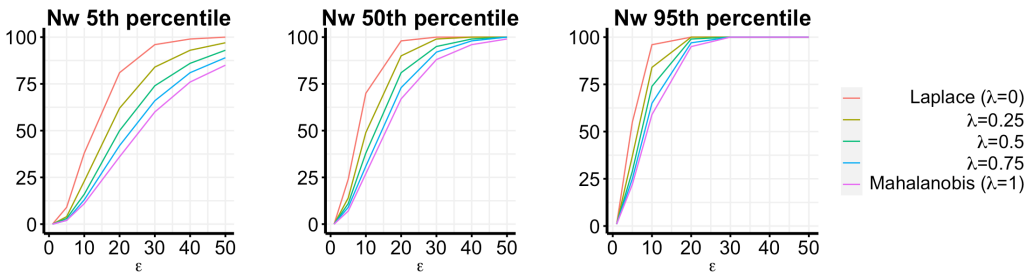

Figure 2: Percentiles for N_w (number of times an input word w does not change) for 300-d FastText embedding over 100 repetitions. The Mahalanobis mechanism has lower values of N_w than the Laplace mechanism.

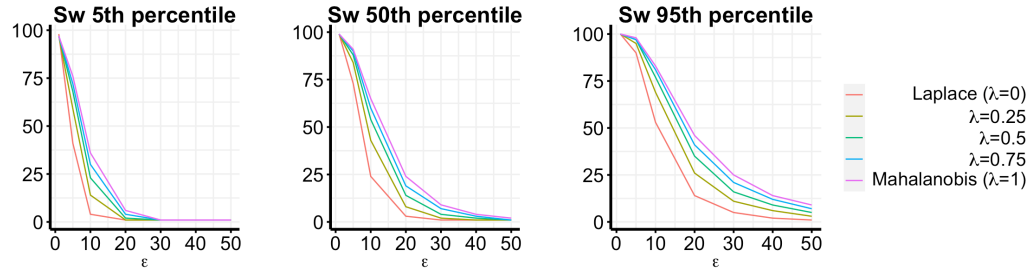

Figure 3: Percentiles for S_w (number of distinct substitutions for an input word w) for 300-d FastText embedding over 100 repetitions. The Mahalanobis mechanism has higher values of S_w than the Laplace mechanism.

ϵ	1	5	10	20	40
Laplace ($\lambda = 0$)	0.19±0.46	26.89±14.49	68.93±17.97	95.28±6.46	99.85±0.68
$\lambda = 0.25$	0.15±0.41	16.66±10.53	50.31±18.68	86.34±12.41	98.69±2.81
$\lambda = 0.5$	0.13±0.38	12.68±8.60	40.40±17.41	78.44±15.40	96.91±4.88
$\lambda = 0.75$	0.12±0.36	10.46±7.49	34.14±16.03	71.81±17.00	94.85±6.79
Mahalanobis ($\lambda = 1$)	0.10±0.35	8.97±6.74	29.73±14.83	66.24±17.90	92.80±8.31

Table 1: Mean ± Standard Deviation for N_w for 300-d FastText embedding. For $\epsilon = 5, 10, 20$, the Mahalanobis mechanism has significantly lower values of mean N_w than the Laplace mechanism, where statistical significance is established by comparing the 95% confidence intervals in the form of $mean \pm 1.96 \times std/\sqrt{100}$.

ϵ	1	5	10	20	40
Laplace ($\lambda = 0$)	99.23±0.87	70.27±15.07	25.82±15.06	4.53±4.56	1.13±0.55
$\lambda = 0.25$	99.18±0.90	81.42±11.12	42.59±16.56	10.31±8.08	1.93±1.80
$\lambda = 0.5$	99.07±0.96	85.69±9.10	52.26±15.96	15.66±10.18	2.98±2.83
$\lambda = 0.75$	98.95±1.03	88.08±7.96	58.66±15.13	20.38±11.46	4.16±3.75
Mahalanobis ($\lambda = 1$)	98.88±1.04	89.58±7.17	63.28±14.25	24.60±12.37	5.32±4.51

Table 2: Mean ± Standard Deviation for S_w for 300-d FastText embedding. For $\epsilon = 5, 10, 20$, the Mahalanobis mechanism has significantly higher values of mean S_w than the Laplace mechanism, where statistical significance is established by comparing the 95% confidence intervals in the form of $mean \pm 1.96 \times std/\sqrt{100}$.

the statistical significance is established by comparing the 95% confidence intervals for mean N_w and S_w in the form of $mean \pm 1.96 \times std/\sqrt{100}$. While the scale of the noise is controlled to be the same across settings, the probability that a word does not change becomes smaller and the number of distinct substitutions becomes larger in our proposed mechanism. This shows the advantage of the Mahalanobis mechanism over the multivariate Laplace mechanism in privacy statistics.

The results are qualitatively similar when we repeat the same set of privacy experiments using the 300-d GloVe embeddings (Pennington et al., 2014). As can be seen in Figure 4, Figure 5, Table 3, and Table 4), the Mahalanobis mechanism has lower values of N_w and higher values of S_w compared

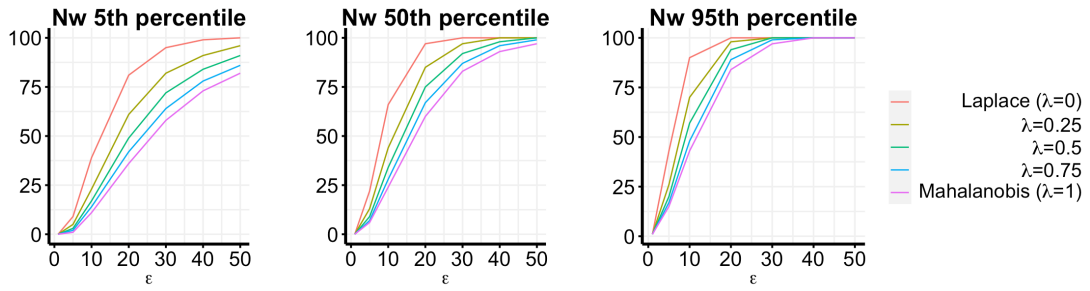

Figure 4: Percentiles for N_w (number of times an input word w does change) for 300-d GloVe embedding over 100 repetitions. The Mahalanobis mechanism has lower values of N_w than the Laplace mechanism.

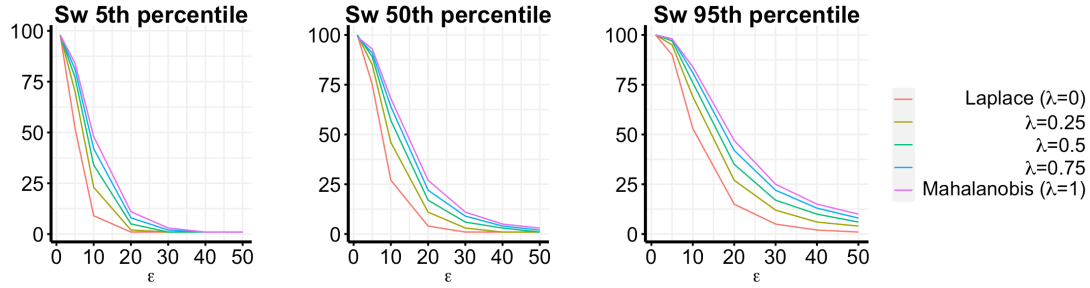

Figure 5: Percentiles for S_w (number of distinct substitutions for an input word w) for 300-d GloVe embedding over 100 repetitions. The Mahalanobis mechanism has higher values of S_w than the Laplace mechanism.

ϵ	1	5	10	20	40
Laplace ($\lambda = 0$)	0.16±0.41	23.71±10.69	65.29±15.53	94.49±6.48	99.81±0.74
$\lambda = 0.25$	0.11±0.34	13.82±6.89	44.83±14.17	83.17±11.61	98.22±3.31
$\lambda = 0.5$	0.09±0.31	10.20±5.57	34.91±12.12	73.78±13.59	95.85±5.56
$\lambda = 0.75$	0.08±0.29	8.16±4.87	28.97±10.72	66.29±14.35	93.26±7.30
Mahalanobis ($\lambda = 1$)	0.08±0.28	6.81±4.36	24.90±9.60	60.27±14.54	90.60±8.65

Table 3: Mean ± Standard Deviation for N_w for 300-d GloVe embedding. For $\epsilon = 5, 10, 20$, the Mahalanobis mechanism has significantly lower values of mean N_w than the Laplace mechanism, where statistical significance is established by comparing the 95% confidence intervals in the form of $mean \pm 1.96 \times std/\sqrt{100}$.

ϵ	1	5	10	20	40
Laplace ($\lambda = 0$)	99.46±0.74	73.20±11.60	28.56±13.61	5.18±4.64	1.16±0.62
$\lambda = 0.25$	99.45±0.73	84.30±7.62	46.08±13.98	12.12±7.69	2.21±2.02
$\lambda = 0.5$	99.38±0.79	88.39±6.08	56.07±12.92	17.96±9.33	3.55±3.06
$\lambda = 0.75$	99.30±0.83	90.71±5.27	62.69±11.84	23.11±10.42	4.94±3.90
Mahalanobis ($\lambda = 1$)	99.24±0.86	92.17±4.65	67.45±10.86	27.51±11.12	6.39±4.57

Table 4: Mean ± Standard Deviation for S_w for 300-d GloVe embedding. For $\epsilon = 5, 10, 20$, the Mahalanobis mechanism has significantly higher values of mean S_w than the Laplace mechanism, where statistical significance is established by comparing the 95% confidence intervals in the form of $mean \pm 1.96 \times std/\sqrt{100}$.

to the Laplace mechanism, which demonstrates a better privacy guarantee.

5.2 Utility Experiments

In the utility experiments, we compare the Mahalanobis mechanism with the multivariate Laplace mechanism in terms of text classification performance on the two real datasets.

On the Twitter Dataset, the task is to classify whether a tweet describes a disaster event, where the benchmark FastText model (Joulin et al., 2016) achieves 0.78 accuracy, 0.78 precision, and 0.69 recall. On the SMSSpam Dataset, the task is spam classification, where the benchmark Bag-of-Words

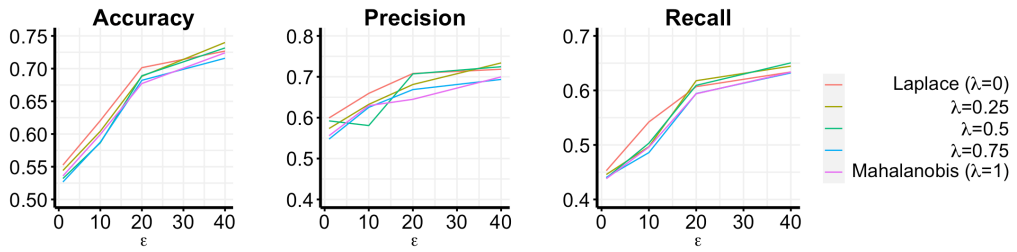

Figure 6: Text classification results on Twitter Dataset. There is no significant difference across mechanisms in terms of accuracy, precision, and recall. This shows the utility is maintained at the same level in the proposed Mahalanobis mechanism across λ.

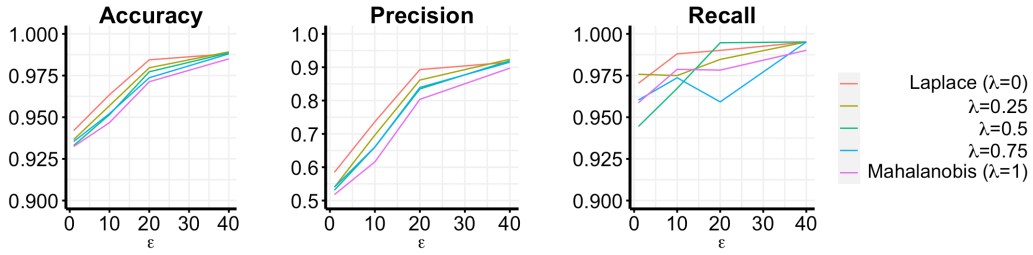

Figure 7: Text classification results on SMSSpam Dataset. There is no significant difference across mechanisms in terms of accuracy, precision, and recall. This shows the utility is maintained at the same level in the proposed Mahalanobis mechanism across λ.

model achieves 0.99 accuracy, 0.92 precision, and 0.99 recall.

In both tasks, we use 70% of the data for training, and 30% of the data for testing. The word embedding vectors are from 300-d FastText. Figure 6 and Figure 7 present the utility results in terms of accuracy, precision and recall on two text classification tasks, respectively. As a general trend in both Twitter and SMSSpam Dataset, the classification accuracy increases with ϵ and eventually approaches the benchmark performance, which is as expected. There are cases where the recall drops when ϵ increases in SMSSpam Dataset, but such drop is not significant as the recall values are all around 0.95 or higher. Across the range of λ, the difference between utility is negligible between the Mahalanobis mechanism and the multivariate Laplace mechanism. Together with results in Section 5.1, we conclude that our proposed mechanism improves the privacy statistics while maintaining the utility at the same level.

6 Conclusions

We develop a differentially private Mahalanobis mechanism for text perturbation. Compared to the existing multivariate Laplace mechanism, our mechanism exploits the geometric property of elliptical noise so as to improve the privacy statistics while maintaining a similar level of utility. Our method can be readily extended to the privacy-preserving analysis on other natural language processing tasks, where utility can be defined according to specific needs.

We remark that the choice of Σ as the global covariance matrix of the word embeddings can be generalized to the personalized covariance matrix within the neighborhood of each word. In this sense, local sensitivity can be used instead of global sensitivity to calibrate the privacy-utility tradeoff. This can be done by adding a preprocessing clustering step on the word embeddings in the vocabulary, and then perform the Mahalanobis mechanism within each cluster using the cluster-specific covariance matrix.

Furthermore, the choice of the tuning parameter λ can also be formulated as an optimization problem with respect to pre-specified privacy and utility constraints. Since λ is the only tuning parameter on a bounded interval of $[0, 1]$, a grid search would suffice, which can be conducted by finding the λ value that maximizes the utility (privacy) objective given the fixed privacy (utility) constraints.

References

Tiago A Almeida, José María G Hidalgo, and Akebo Yamakami. 2011. Contributions to the study of sms

spam filtering: new collection and results. In *Proceedings of the 11th ACM symposium on Document engineering*, pages 259–262.

Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. 2012. t-plausibility: Generalizing words to desensitize text. *Trans. Data Priv.*, 5(3):505–534.

Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914.

Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019. I am not what i write: Privacy preserving text representation learning. *arXiv preprint arXiv:1907.03189*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer.

Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. 2015. Constructing elastic distinguishability metrics for location privacy. *Proceedings on Privacy Enhancing Technologies*, 2015(2):156–170.

Romain Couillet and Matthew McKay. 2014. Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, 131:99–120.

Chad Cumby and Rayid Ghani. 2011. A machine learning based system for semi-automatically redacting documents. In *Twenty-Third IAAI Conference*.

Michael J Daniels and Robert E Kass. 2001. Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184.

Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18.

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

C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan. 2015. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669.

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, Cham.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 178–186.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. 2013. Differential privacy for functions and functional data. *Journal of Machine Learning Research*, 14(Feb):703–727.

Rob Hall et al. 2012. *New Statistical Applications for Differential Privacy*. Ph.D. thesis, PhD thesis, PhD thesis, Carnegie Mellon.

Steven Hill, Zhimin Zhou, Lawrence Saul, and Hovav Shacham. 2016. On the (in) effectiveness of mosaicing and blurring as tools for document redaction. *Proceedings on Privacy Enhancing Technologies*, 2016(4):403–417.

et al. Homer N, Szelinger S. 2008. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8)(e1000167).

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826.

Prasanta Chandra Mahalanobis. 1936. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55.

A. Narayanan and V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125.

Rina Okada, Kazuto Fukuchi, and Jun Sakuma. 2015. Differentially private analysis of outliers. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 458–473. Springer.

Hwee Hwa Pang, Xuhua Ding, and Xiaokui Xiao. 2010. Embellishing text search queries to protect user privacy.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Albin Petit, Thomas Cerqueus, Sonia Ben Mokhtar, Lionel Brunie, and Harald Kosch. 2015. Peas: Private, efficient and accurate web search. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 571–580. IEEE.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

David SáNchez, Jordi Castellà-Roca, and Alexandre Viejo. 2013. Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines. *Information Sciences*, 218:17–30.

et al. Sankararaman S., Obozinski G. 2009. Genomic privacy and limits of individual detection in a pool. *Nat Genet*, 41:965–967.

Juliane Schäfer and Korbinian Strimmer. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1).

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Latanya Sweeney. 1997. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110. PMID: 11066504.

Isabel Wagner and David Eckhoff. 2018. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):1–38.

Rik Warren, Robert F Smith, and Anne K Cybenko. 2011. Use of mahalanobis distance for detecting outliers and outlier clusters in markedly non-normal data: a vehicular traffic example. Technical report, SRA INTERNATIONAL INC DAYTON OH.

Shiming Xiang, Feiping Nie, and Changshui Zhang. 2008. Learning a mahalanobis distance metric for data clustering and classification. *Pattern recognition*, 41(12):3600–3612.

Yuxiang Zhang, Bo Du, Liangpei Zhang, and Shugen Wang. 2015. A low-rank and sparse matrix decomposition-based mahalanobis distance method for hyperspectral anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1376–1389.

Xuemei Zhao, Yu Li, and Quanhua Zhao. 2015. Mahalanobis distance based on fuzzy clustering algorithm for image segmentation. *Digital Signal Processing*, 43:8–16.

A Additional Information on the Twitter and SMSSpam Data

Data	Twitter	SMSSpam
Number of records	7,613	5,574
Words per record	7.3±3.0	15.6±11.4
Vocabulary size	22,213	9,515
d_{max}/d_{min}	35.28	8.92
$\overline{d}_{:50}/\overline{d}_{-50:}$	4.47	5.67
$\overline{p}_{:50}$	0.1%	0.4%

Table 5: d_{max} is the largest Euclidean distance to the nearest word embedding in the vocabulary; d_{min} is the smallest Euclidean distance to the nearest word embedding in the vocabulary; $\overline{d}_{:50}$ is the mean distance to the nearest word embedding for the top 50 words in the sparse regions (largest distance to nearest neighbors); $\overline{d}_{-50:}$ is the mean distance to the nearest word embedding for the top 50 words in the dense regions (smallest distance to nearest neighbors). These ratio statistics demonstrate the heterogeneity in the density of the word embedding space. $\overline{p}_{:50}$ is the mean percentage of records that contains the top 50 words in the sparse regions, which suggests that those words in the sparse regions are indeed rare words that can be used to link to specific records.