Calibrating Mechanisms for Privacy Preserving Text Analysis

Oluwaseyi Feyisetan
Amazon
sey@amazon.com

Borja Balle
Deep Mind
borja.balle@gmail.com

Tom Diethe
Amazon
tdiethe@amazon.com

Thomas Drake
Amazon
draket@amazon.com

ABSTRACT

This talk presents a formal approach to carrying out privacy preserving text perturbation using a variant of Differential Privacy (DP) known as Metric DP (mDP). Our approach applies carefully calibrated noise to vector representation of words in a high dimension space as defined by word embedding models. We present a privacy proof that satisfies mDP where the privacy parameter ε provides guarantees with respect to a distance metric defined by the word embedding space. We demonstrate how ε can be selected by analyzing plausible deniability statistics backed up by large scale analysis on GloVe and fastText embeddings. We also conduct experiments on well-known datasets to demonstrate the tradeoff between privacy and utility for varying values of ε on different task types. Our results provide insights into carrying out practical privatization on text-based applications for a broad range of tasks.

CCS CONCEPTS

• **Security and privacy → Privacy protections**;

ACM Reference Format:
Oluwaseyi Feyisetan, Borja Balle, Tom Diethe, and Thomas Drake. 2020. Calibrating Mechanisms for Privacy Preserving Text Analysis. In *Proceedings of Workshop on Privacy in Natural Language Processing (PrivateNLP '20).* Houston, TX, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 METRIC DIFFERENTIAL PRIVACY

Over the last decade, Differential Privacy (DP) [5] has emerged as a *de facto* standard for privacy-preserving data analysis algorithms. One reason for such success is the robustness of DP against critical pitfalls exhibited by previous attempts to formalize privacy in the context of data analysis algorithms. Several variants of DP have been proposed in the literature to address a variety of settings depending on whether, for example, privacy is defined with respect to aggregate statistics and Machine Learning (ML) models (*curator DP*) [5], or privacy is defined with respect to the data points contributed by each individual (*local DP*) [8].

Since our application involves privatizing individual sentences submitted by each user, Local DP (LDP) would be the ideal privacy model to consider. However, LDP has a requirement that renders it impractical for our application: it requires that the secret sentence x_s has a non-negligible probability of being transformed into *any* other sentence $\hat{x}_s$, no matter how unrelated x_s and $\hat{x}_s$ are. Unfortunately, this constraint makes it virtually impossible to enforce that the *semantics* of x_s are approximately captured by the privatized sentence $\hat{x}_s$, since the space of sentences is exponentially large in the length $|x_s|$, and the number of sentences semantically related to x_s will have vanishingly small probability under LDP.

To address this limitation we adopt *metric DP* (mDP) [1, 4], a relaxation of local DP that originated in the context of location privacy to address precisely the limitation described above. In particular, mDP allows a mechanism to report a user's location in a privacy-preserving manner, while giving higher probability to locations which are close to the current location, and negligible probability to locations in a completely different part of the planet. Metric DP was originally developed as an abstraction of the privacy model proposed in [2] to address the privacy-utility trade-off in location privacy. To the best of our knowledge, our paper [6, 7] was the first to use mDP in the context of language data.

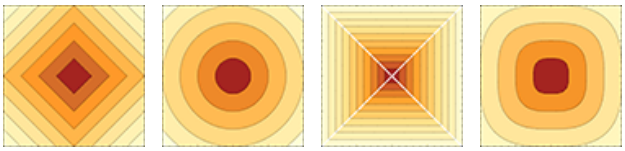

Figure 1: Contour plots of different metrics [11]. Left to right: L_1 Manhattan distance, L_2 Euclidean distance, L_∞ Chebyshev distance, L_p Minkowski distance (L_3 shown here)

Formally, mDP is defined for mechanisms whose inputs come from a set $\mathcal{X}$ equipped with a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ satisfying the axioms of a metric (*i.e.* identity of indiscernibles, symmetry and triangle inequality). The definition of mDP depends on the particular distance function d being used and it is parametrized by a privacy parameter $\varepsilon > 0$. We say that a randomized mechanism $M : \mathcal{X} \to \mathcal{Y}$ satisfies ε-mDP if for any $x, x' \in \mathcal{X}$ the distributions over outputs of $M(x)$ and $M(x')$ satisfy the following bound: for all $y \in \mathcal{Y}$ we have

$$\frac{\Pr[M(x) = y]}{\Pr[M(x') = y]} \leq e^{\varepsilon d(x,x')} \quad . \tag{1}$$

We note that mDP exhibits the same desirable properties of DP (*e.g.* composition, post-processing, robustness against side knowledge, *etc.*), but we shall not be using these properties explicitly in our analysis; we refer the reader to [4] for further details.

The type of probabilistic guarantee described by (1) is characteristic of DP: it says that the log-likelihood ratio of observing any

PrivateNLP '20, February 7, 2020, Houston, TX, USA

particular output y given two possible inputs x and x' is bounded by $\varepsilon d(x, x')$. The key difference between mDP and local DP is that the latter corresponds to a particular instance of the former when the distance function is given by $d(x, x') = 1$ for every $x \neq x'$. Unfortunately, this Hamming metric does not provide a way to classify some pairs of points in $\mathcal{X}$ as being closer than others. This indicates that local DP implies a strong notion of indistinguishability of the input, thus providing very strong privacy by "remembering almost nothing" about the input. In contrast, metric DP is less restrictive and allows the indistinguishability of the output distributions to be scaled by the distance between the respective inputs. In particular, the further away a pair of inputs are, the more distinguishable the output distributions can be, thus allowing these distributions to remember more about their inputs than under the strictly stronger definition of local DP.

A point of consideration for mDP is that the meaning of the privacy parameter ε changes if one considers different metrics, and is in general incomparable with the ε parameter used in standard (local) DP. Thus, in order to understand the privacy consequences of a given ε in mDP one needs to understand the structure of the underlying metric d.

2 PRIVACY MECHANISM

Our privacy mechanism is described over the metric space induced by word embeddings as follows:

Algorithm 1: Privacy Preserving Mechanism

Input: string $x = w_1 w_2 \cdots w_\ell$, privacy parameter $\varepsilon > 0$
for $i \in \{1, \ldots, \ell\}$ **do**
  Compute embedding $\phi_i = \phi(w_i)$
  Perturb embedding to obtain $\hat{\phi}_i = \phi_i + N$ with noise density $p_N(z) \propto \exp(-\varepsilon \|z\|)$
  Obtain perturbed word $\hat{w}_i = argmin_{u \in \mathcal{W}} \|\phi(u) - \hat{\phi}_i\|$
  Insert $\hat{w}_i$ in ith position of $\hat{x}$
release $\hat{x}$

3 STATISTICS FOR PRIVACY CALIBRATION

We now present a methodology for calibrating the ε parameter of our mDP mechanism $\mathcal{M}$ based on the geometric structure of the word embedding ϕ used to define the metric d. Our strategy boils down to identifying a small number of statistics associated with the output distributions of $\mathcal{M}$, and finding a range of parameters ε where these statistics behave as one would expect from a mechanism providing a prescribed level of plausible deniability. We recall that the main reason this is necessary, and why the usual rules of thumb for calibrating ε in traditional (*i.e.* non-metric) DP cannot be applied here, is because the meaning of ε in mDP depends on the particular metric being used and is not transferable across metrics. We start by making some qualitative observations about how ε affects the behavior of mechanism $\mathcal{M}$. For the sake of simplicity we focus the discussion on the case where x is a single word $x = w$, but all our observations can be directly generalized to the case $|x| > 1$. We note these observations are essentially heuristic, although it is not hard to turn them into precise mathematical statements.

A key difference between LDP and mDP is that the former provides a stronger form of plausible deniability by insisting that almost every outcome is possible when a word is perturbed, while the later only requires that we give enough probability mass to words close to the original one to ensure that the output does not reveal what the original word was, although it still releases information about the neighborhood where the original word was.

More formally, the statistics we look at are the probability $N_w = \Pr[M(w) = w]$ of not modifying the input word w, and the (effective) support of the output distribution S_w (*i.e.* number of possible output words) for an input w. In particular, given a small probability parameter $\eta > 0$, we define S_w as the size of the smallest set of words that accumulates probability at least $1 - \eta$ on input w:

$$S_w = \min |\{S \subseteq \mathcal{X} \ : \ \Pr[M(w) \notin S] \leq \eta\}| \ .$$

Intuitively, a setting of ε providing plausible deniability should have N_w small and S_w large for (almost) all words in $w \in \mathcal{W}$.

These statistics can also be related to the two extremes of the Rényi entropy [10], thus providing an additional information-theoretic justification for the settings of ε that provide plausible deniability in terms of large entropy. Recall that for a distribution p over $\mathcal{W}$ with $p_w = \Pr_{W \sim p}[W = w]$, the Rényi entropy of order $\alpha \geq 0$ is given by

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \left(\sum_{w \in \mathcal{W}} p_w^\alpha \right) \ .$$

The Hartley entropy H_0 is the special case of Rényi entropy with $\alpha = 0$. It depends on vocabulary size $|\mathcal{W}|$ and is therefore a best-case scenario as it represents the perfect privacy scenario for a user as the number of words grow. It is given by $H_0 = \log_2 |\mathcal{W}|$. Min-entropy H_∞ is the special case with $\alpha = \infty$ which is a worst-case scenario because it depends on the adversary attaching the highest probability to a specific word $p(w)$. It is given by $H_\infty = -\log_2 \max_{w \in \mathcal{W}} (p(w))$.

This implies that we can see the quantities S_w and N_w as proxies for the two extreme Rényi entropies through the approximate identities $H_0(\mathcal{M}(w)) \approx \log S_w$ and $H_\infty(\mathcal{M}(w)) \approx \log 1/N_w$, where the last approximation relies on the fact that (at least for small enough ε), w should be the most likely word under the distribution of $\mathcal{M}(w)$.

4 WORD EMBEDDINGS

A word embedding $\phi : \mathcal{W} \to \mathbb{R}^n$ maps each word in some vocabulary to a vector of real numbers.

The geometry of the resulting embedding model has a direct impact on defining the output distribution of our redaction mechanism. To get an intuition for the structure of these metric spaces – *i.e.*, how words cluster together and the distances between words and their neighbors – we ran several analytical experiments on two widely available word embedding models: GloVe [9] and fastText [3]. We selected 319,000 words that were present in both the GloVe and fastText embeddings. Though we present findings only from the common 319,000 words in the embedding vocabularies, we carried out experiments over the entire vector space (*i.e.*, 400,000 for GloVe and 2,519,370 for fastText).

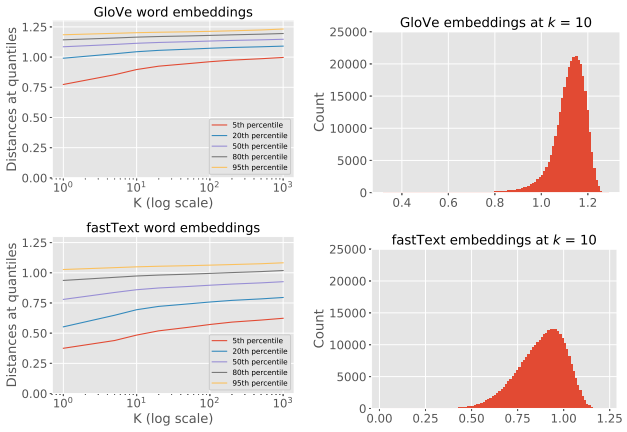

Figure 2: Distribution of distances between a given vector and its k closest neighbors for GloVe and fastText

Our experiments provide: (i) insights into the distance $d(x, x')$ that controls the privacy guarantees of our mechanism for different embedding models; and (ii) empirical evaluation of the plausible deniability statistics S_w and N_w for the mechanisms obtained using different embeddings.

5 CHOOSING A WORD EMBEDDING MODEL

Our analysis gives a reasonable approach to selecting ε by means of the proxies provided by the plausible deniability statistics. In general, tuning privacy parameters in (metric) differential privacy is still a topic under active research, especially with respect to what ε means for different applications. Task owners ultimately need to determine what is best for their users based on the available descriptive statistics.

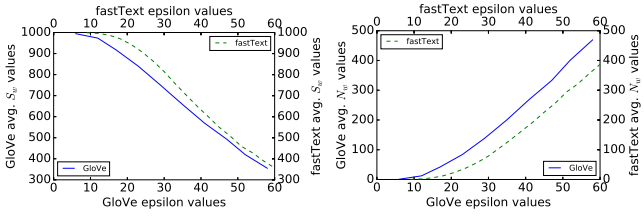

Figure 3: Average S_w and N_w statistics for GloVe and fastText

In Fig. 3, we present the average values of S_w and N_w statistics for GloVe and fastText. We observe that similar values over fastText cover a broader range of ε values when compared to GloVe. However, the average values are not sufficient to make a conclusive comparison between embedding models. The shape of the distribution needs to be further considered when selecting ε for similar values of N_w *or* S_w. For example, the average value of S_w for GloVe at $\varepsilon = 29$ is about the same as that for fastText at $\varepsilon = 33$ (both have an average S_w value of ≈ 750). However, from Fig. 4, we discover that they both have slightly different distributions with GloVe being slightly flatter and fastText more skewed to the right. Similar results for matching average values of N_w for GloVe at $\varepsilon = 29$ and fastText at $\varepsilon = 36$ are also presented in Fig. 4.

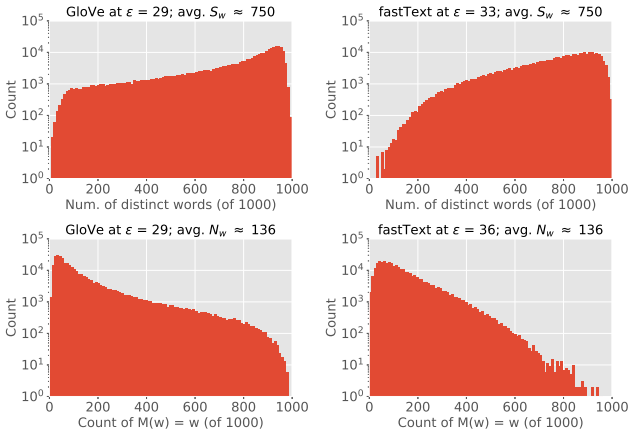

Figure 4: Comparing the distribution of similar average S_w and N_w results for GloVe and fastText

6 MORE CALIBRATION STATISTICS

In this section, we highlight the results of empirical S_w and N_w plausible deniability statistics for 300 dimension fastText and 50 dimension GloVe embeddings. The results were designed to yield comparable numbers to those presented for 300 dimension GloVe embeddings.

7 BIOGRAPHY

Dr. Oluwaseyi Feyisetan is an Applied Scientist at Amazon Alexa where he works on Differential Privacy and Privacy Auditing mechanisms within the context of Natural Language Processing. He holds 2 pending patents with Amazon on preserving privacy in NLP systems. He completed his PhD at the University of Southampton in the UK and has published in top tier conferences and journals on crowdsourcing, homomorphic encryption, and privacy in the context of Active Learning and NLP. He has served as a reviewer at top NLP conferences including ACL and EMNLP. He is the lead organizer of the Workshop on Privacy and Natural Language Processing (PrivateNLP) at WSDM with an upcoming event scheduled for EMNLP. Prior to working at Amazon in the US, he spent 7 years in the UK where he worked at different startups and institutions focusing on regulatory compliance, machine learning and NLP within the finance sector, most recently, at the Bank of America.

REFERENCES

[1] Mário Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. 2018. Local Differential Privacy on Metric Spaces: optimizing the trade-off with utility. In *Computer Security Foundations Symposium (CSF)*.

[2] 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 CCS*. ACM, 901–914.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *TACL* 5 (2017).

[4] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *Intl. Symposium on Privacy Enhancing Technologies Symposium*.

Figure 5: Empirical S_w and N_w statistics for fastText word embeddings as a function of ε.

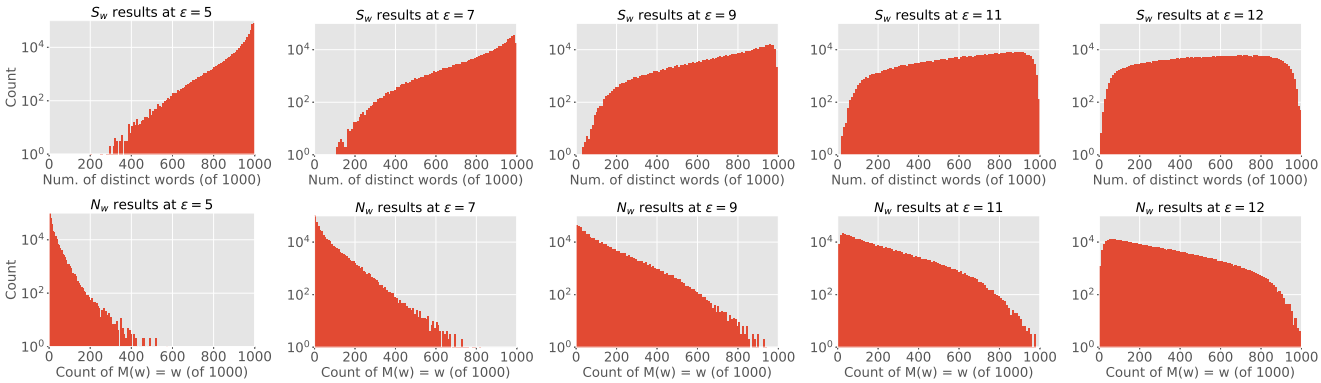

Figure 6: Empirical S_w and N_w statistics for 50 dimensional GloVe word embeddings as a function of ε.

[5] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC*. Springer, 265–284.

[6] 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*. 178–186.

[7] Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging Hierarchical Representations for Preserving Privacy and Utility in Text. In *IEEE International Conference on Data Mining (ICDM)*.

[8] Shiva Kasiviswanathan, Homin Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM J. Comput.* 40, 3 (2011).

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[10] Alfréd Rényi. 1961. *On measures of entropy and information.* Technical Report. HUNGARIAN ACADEMY OF SCIENCES Budapest Hungary.

[11] Christoph Ruegg, Marcus Cuda, and Jurgen Van Gael. 2009. Distance Metrics. (2009). https://numerics.mathdotnet.com/Distance.html