Privacy-Preserving Textual Analysis via Calibrated Perturbations

Oluwaseyi Feyisetan
Amazon
sey@amazon.com

Borja Balle
Amazon
pigem@amazon.co.uk

Thomas Drake
Amazon
draket@amazon.com

Tom Diethe
Amazon
tdiethe@amazon.co.uk

ABSTRACT

Accurately learning from user data while providing quantifiable privacy guarantees provides an opportunity to build better ML models while maintaining user trust. This paper presents a formal approach to carrying out privacy preserving text perturbation using the notion of d_χ-privacy designed to achieve geo-indistinguishability in location data. 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 d_χ-privacy 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 conduct privacy audit experiments against 2 baseline models and utility experiments on 3 datasets to demonstrate the tradeoff between privacy and utility for varying values of ε on different task types. Our results demonstrate practical utility (< 2% utility loss for training binary classifiers) while providing better privacy guarantees than baseline models.

ACM Reference Format:
Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-Preserving Textual Analysis via Calibrated Perturbations. In *Proceedings of Workshop on Privacy and Natural Language Processing (PrivateNLP '20).* Houston, TX, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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

Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations

Oluwaseyi Feyisetan sey@amazon.com

Borja Balle borja.balle@gmail.com

Thomas Drake draket@amazon.com

Tom Diethe tdiethe@amazon.com

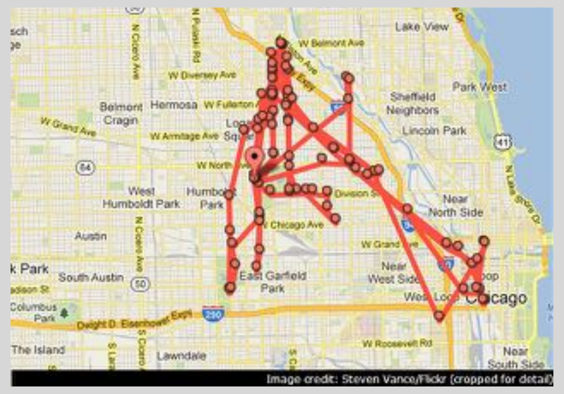
Summary

- **User's goal:** meet some specific need with respect to an issued query *x*
- **Agent's goal:** satisfy the user's request
- **Question:** what occurs when *x* is used to make other inferences about the user

- **Mechanism:** modify the query to protect privacy whilst preserving semantics
- **Our approach:** Generalized Metric Differential Privacy.

Introduction

What makes privacy difficult?



High dimensional data

Big and richer datasets lead to users generating uniquely identifiable information.

Side knowledge

Innocuous data reveals customer information when joined with side-knowledge.

Privacy in textual data


A Face Is Exposed for AOL Searcher No. 4417749

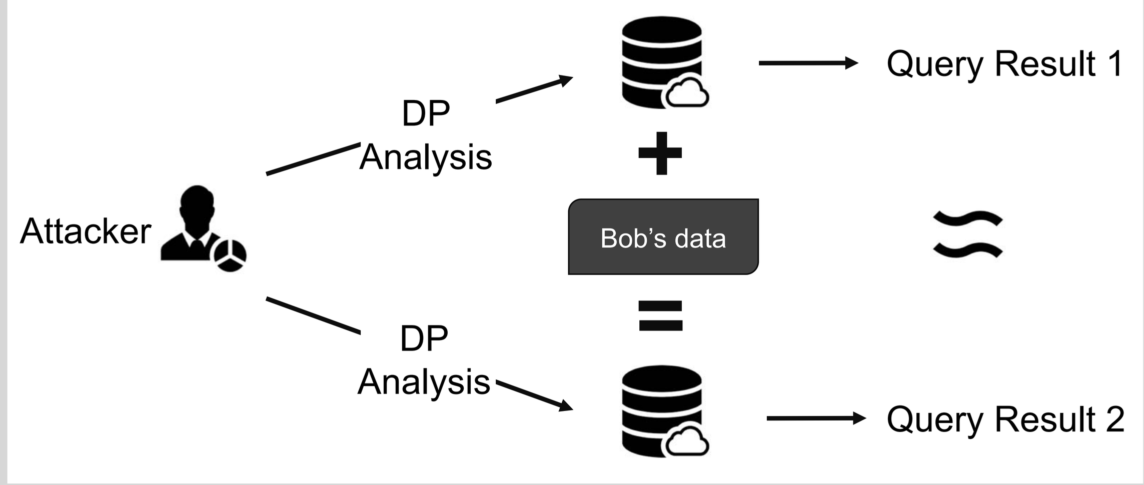
By MICHAEL BARBARO and TOM ZELLER Jr. AUG. 9, 2006 NEW YORK TIMES ARTICLE

User	Text
441779	dog that urinates on everything
441779	safest place to live
. . .	
441779	the best season to visit Italy
441779	landscapers in Lilburn, GA

Most of the queries do not contain PII

A viable solution: Differential Privacy

ε-Differential Privacy (DP) bounds the influence of any single input on the output of a computation.



Result 1 is approximately equal to Result 2

Differential Privacy

A *randomized* mechanism $\mathcal{M} : X \mapsto Y$ is ε-differentially private if for all neighboring inputs $x \simeq x'$ (i.e., $d_h(x, x') = 1$ where d_h is the Hamming distance) and for all set of outputs $E \subseteq Y$,

$$\mathbb{P}[\mathcal{M}(x) \in E] \leq e^{\varepsilon d_h(x,x')} \mathbb{P}\left[\mathcal{M}(x') \in E\right]$$

Metric DP generalizes this to use any valid metric $d_h(x, x')$, (i.e., one that satisfies non negativity, indiscernibles, symmetry, and triangle inequality)

Generalized Metric Differential Privacy

Metric DP is a parameterized by a distance measure *d* and, conceptually, increases the area where a person's location probably lies

Represent using word embeddings which map words into a vector space $\phi : w \mapsto \mathbb{R}^n$

Mechanism Overview

We sample noise from the multivariate Laplacian distribution to achieve ε –mDP

- **Robust to post-processing**

If $\mathcal{M}$ is ε-DP, then $f(\mathcal{M})$ is at least ε-DP

- **Composition**

If $\mathcal{M}_1, \ldots, \mathcal{M}_n$ are ε-DP, $g(\mathcal{M}_1, \ldots, \mathcal{M}_n)$ is $\sum_{i=1}^{n} \varepsilon_i$-DP by additive composition

- **Protects against side knowledge**

If attacker has prior p_1 and computes posterior p_2 after observing output of ε-DP, then $dist(p_1, p_2) = \mathcal{O}(\varepsilon)$

Mechanism Details

Inputs:

- $w \in W$: word to be 'privatized'
- $\phi : W \mapsto Z$: embedding function
- $d : Z \times Z \mapsto \mathbb{R}$: distance function
- $\Omega(\varepsilon)$: DP noise distribution

1. Project word $v = \phi(w)$
2. Perturb $v' = v + \xi$ where $\xi \sim \Omega(\varepsilon)$
3. Vector v' will not be a word (a.s.)
4. Project back to dictionary space $W : w' = \arg\min_{w \in W} d(v', \phi(w))$
5. Return w'

Sampling and Calibration

To sample from the multivariate Laplace distribution: $\Omega(\varepsilon)$

1. Sample a random variable v from the multivariate normal distribution
2. Sample a magnitude l from the Gamma distribution with $^1/_\varepsilon$
3. Return $v.l$

Define statistics to measure the ε privacy:

1. Probability $N_w = \mathrm{P}[\mathcal{M}(w) = w]$ of not modifying input word w and,
2. The (effective) support of the output distribution S_w on $\mathcal{M}(w)$

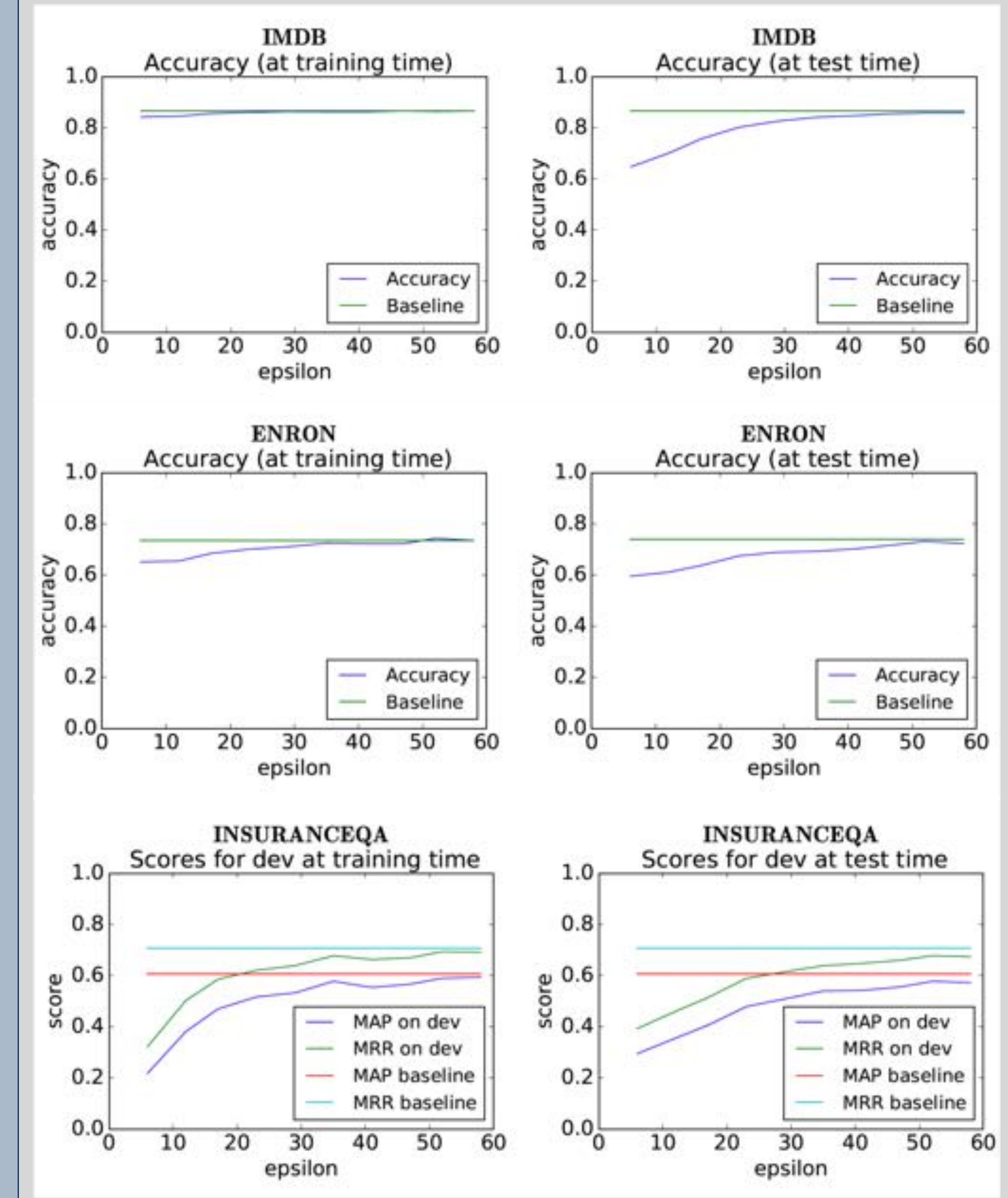
Sample results

ε	Avg. N_w	*w* = encryption: GloVe	*w* = encryption: fastText
⟵ increasing ε, better semantics	50	freebsd multibody 56-bit public-key	ncurses vpns tcp isdn
	100	ciphertexts truecrypt demodulator rootkit	plaintext diffie-hellman multiplexers cryptography
	200	harbormaster unencrypted cryptographically authentication	cryptographic ssl/tls authentication cryptography
	300	decryption encrypt encrypted encryption	encrypt unencrypted encryptions encrypted

Experiment Results

Metric	6	12	17	23	29	35	41	47
Precision	0.00	0.00	0.00	0.00	0.67	0.90	0.93	1.00
Recall	0.00	0.00	0.00	0.00	0.02	0.09	0.14	0.30
Accuracy	0.50	0.50	0.50	0.50	0.51	0.55	0.57	0.65
AUC	0.06	0.04	0.11	0.36	0.61	0.85	0.88	0.93

Scores measure privacy loss (lower is better)



Utility of downstream machine learning model on data (higher is better)