Studying the Privacy-Utility Trade-off of Word Embedding Perturbations with a Focus on Sensitivity Analysis and Vector Mapping

Alisha Riecker

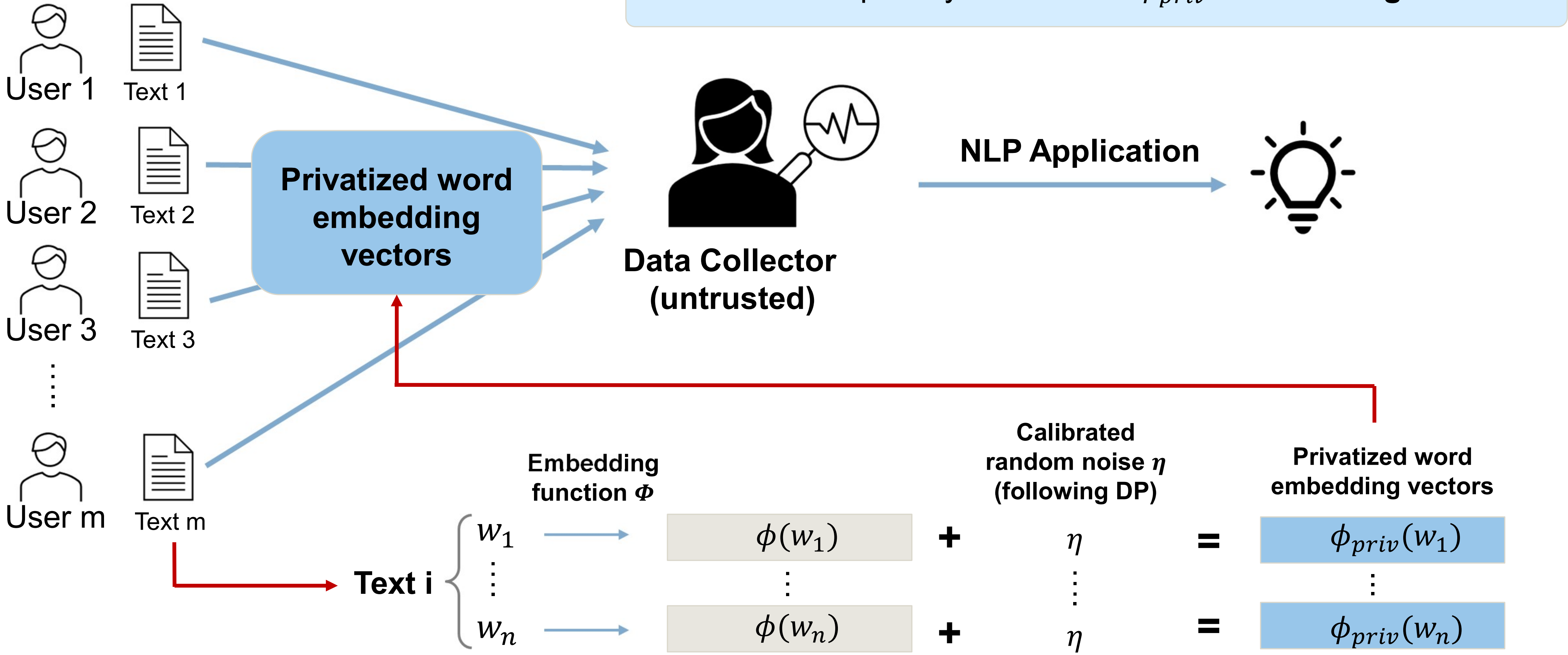
7th August 2023, Master Thesis Kick-off

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

Motivation – Differential Privacy for Word Embedding Vectors (1)

Given any two input words w, w', their corresponding outputs of the differential privacy mechanism ϕ_{priv} are **indistinguishable**.



Motivation - Differential Privacy for Word Embedding Vectors (2)

$\boldsymbol{\Phi(w)}$ (Embedding Function)	+	$\boldsymbol{\eta}$ (Calibrated Random Noise)	=	$\boldsymbol{\Phi_{priv}(w)}$ (Randomized Mechanism)

Example: Laplace Mechanism

An embedding function $\boldsymbol{\Phi(w)}$ with sensitivity Δ_{ϕ}

$\boldsymbol{\eta \sim Lap}$

Φ_{priv} is **$\boldsymbol{\varepsilon}$-differentially private**

*Sensitivity describes the **maximum possible change** in a function's output when the input is changed.*

Variance depends on sensitivity Δ_{ϕ} and privacy budget ε

*The probability of a specific output on any two inputs w, w' can differ by at most a **multiplicative factor** dependent on ε.*

Motivation - Differential Privacy for Word Embedding Vectors (3)

$\boldsymbol{\Phi(w)}$ (Embedding Function) + $\boldsymbol{\eta}$ (Calibrated Random Noise) = $\boldsymbol{\Phi_{priv}(w)}$ (Randomized Mechanism)

Large sensitivity Δ_ϕ leads to large noise η → Bound sensitivity → ?

- How can sensitivity be bounded?
- How does the bounding affect the Privacy-Utility Trade-off?

$\Phi_{priv}(w)$ is not a "real" word → Map $\Phi_{priv}(w)$ to similar embedding vector associated with a "real" word → ?

- What can the mapping look like?
- How does this mapping affect the Privacy-Utility Trade-off?

Outline

1. Motivation
2. Research Questions
3. Methodology
4. Initial Findings
5. Next Steps
6. Timeline

Research Questions

RQ1	What approaches are there to privatize word embeddings by **perturbing word vector representations**?
RQ2	How can we make these privatized word embeddings **more effective**?
RQ3	What is the effect of different approaches to **estimating sensitivity** on **privacy and utility** for downstream NLP tasks?
RQ4	What are the implications on **privacy and utility** resulting from **mapping noisy word embeddings** to similar embedding vectors which are associated with real words?

Outline

Methodology – Literature Review

Domain:

- Text data
- Image data
- Location data

Differential Privacy Method:

- Embedding perturbation
- Gradient perturbation

Goals

RQ1

Approaches to privatize word embeddings by **perturbing word vector representations**

RQ2

Ideas to **increase the utility of the privatized word embeddings**
(focus on sensitivity- and mapping-related ideas)

Methodology - Experiments

Empirically compare privatized embedding vectors

Utility Evaluation

- Sentiment Analysis
- Topic Classification

Privacy Evaluation

- Identification of authors' gender and age
- Identification of named entities

Goals

RQ3: Effect of different approaches to **bounding sensitivity** on **privacy and utility**

RQ4: Effect of different approaches to **mapping noisy word embeddings** to similar embedding vectors on **privacy and utility**

Outline

Initial Findings – Bounding Sensitivity

Concatenate Φ with a function f_{bound}

$\Phi(w)$ has unbounded sensitivity ⟶ $\tilde{\Phi}(w) = \boxed{f_{bound} \circ \Phi(w)}$, s.th. $\boxed{\tilde{\Phi}(w) \in [a, b]^d}$

$\tilde{\Phi}(w)$ has bounded sensitivity

Bounding sensitivity using f_{bound}		**Sensitivity**
Bound the embedding space	By observed range	$\Delta_{\tilde{\Phi}} \leq d * 2 * v_{max}$, *where v_{max} is the maximum observed value*
	Normalize to $[0, 1]^d$	$\Delta_{\tilde{\Phi}} \leq d$
	Normalize to unit length	$\Delta_{\tilde{\Phi}} \leq 2$
Dimensionality Reduction	Johnson-Lindenstrauss Lemma	$\Delta_{\tilde{\Phi}} \leq d' * (b - a)$, where $d' < d$

$$f_{bound}(y) = \frac{y - \min(y)}{\max(y) - \min(y)}$$

Initial Findings – Mapping to “real“ words

Concatenate $\Phi_{\boldsymbol{priv}}$ with a function f_{map}

$\Phi_{priv}(w)$ does not correspond to a word embedding vector

$$\Phi_{\boldsymbol{priv\_map}}(w) = \boxed{f_{map} \circ \Phi_{priv}(w)}\text{, s.th. }\boxed{\Phi_{\boldsymbol{priv\_map}}(w) \in Im(\Phi)}$$

$\Phi_{\boldsymbol{priv\_map}}(w)$ corresponds to a “real” word‘s embedding vector

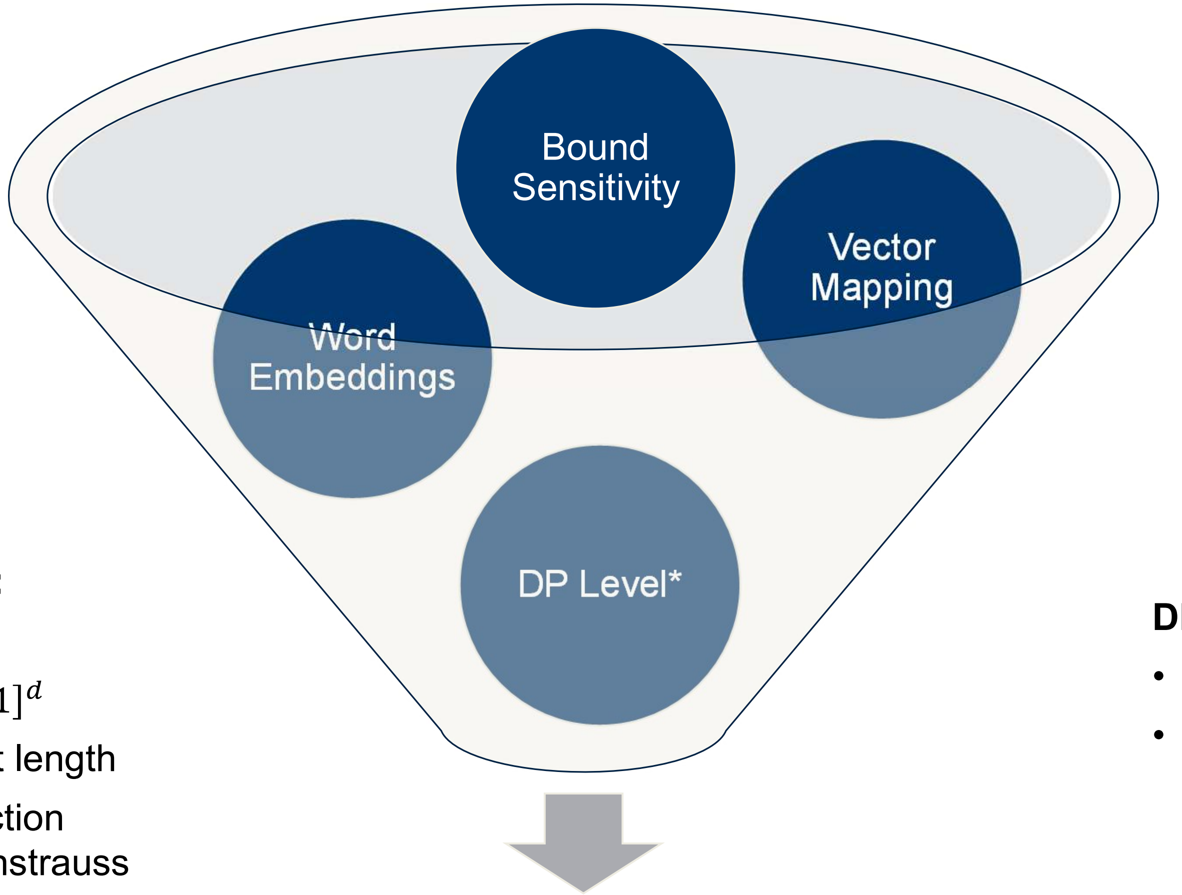
Mapping	Resulting embedding vector
Map to nearest neighbor embedding	Similar embedding vector associated with "real" word
Randomly output first or second nearest neighbor	Similar embedding vector associated with "real" word
No mapping	Noisy embedding vector

Outline

Next Steps - Experiments

Word Embeddings:
- GloVe embeddings
- FastText embeddings

Bound Sensitivity:
- observed range
- normalize to $[0,1]^d$
- normalize to unit length
- dimension reduction (Johnson-Lindenstrauss Lemma)



Compare privatized embedding vectors

Vector Mapping:
- map to nearest neighbor embedding
- randomly output first or second nearest neighbor
- no vector mapping

DP Level:
- Word level privacy
- Document/User level privacy

*Differential Privacy Levels – Document vs. Word Level Privacy

Given any two input texts t, t', their corresponding outputs of the differential privacy mechanism ϕ_{priv} are **ε-indistinguishable**.

Document level privacy

Embedding function Φ

Text i $\begin{cases} w_1 \\ \vdots \\ w_n \end{cases}$ → $\phi(text_i)$ **+** $\eta \sim Lap$ **=** $\phi_{priv}(text_i)$

Calibrated random noise η (e.g., Laplace)

Privatized text embedding vector

Word level privacy

Embedding function Φ

Calibrated random noise η (e.g., Laplace)

Privatized word embedding vectors

Text i

w_1 → $\phi(w_1)$ **+** $\eta \sim Lap$ **=** $\phi_{priv}(w_1)$

$\vdots$

w_n → $\phi(w_n)$ **+** $\eta \sim Lap$ **=** $\phi_{priv}(w_n)$

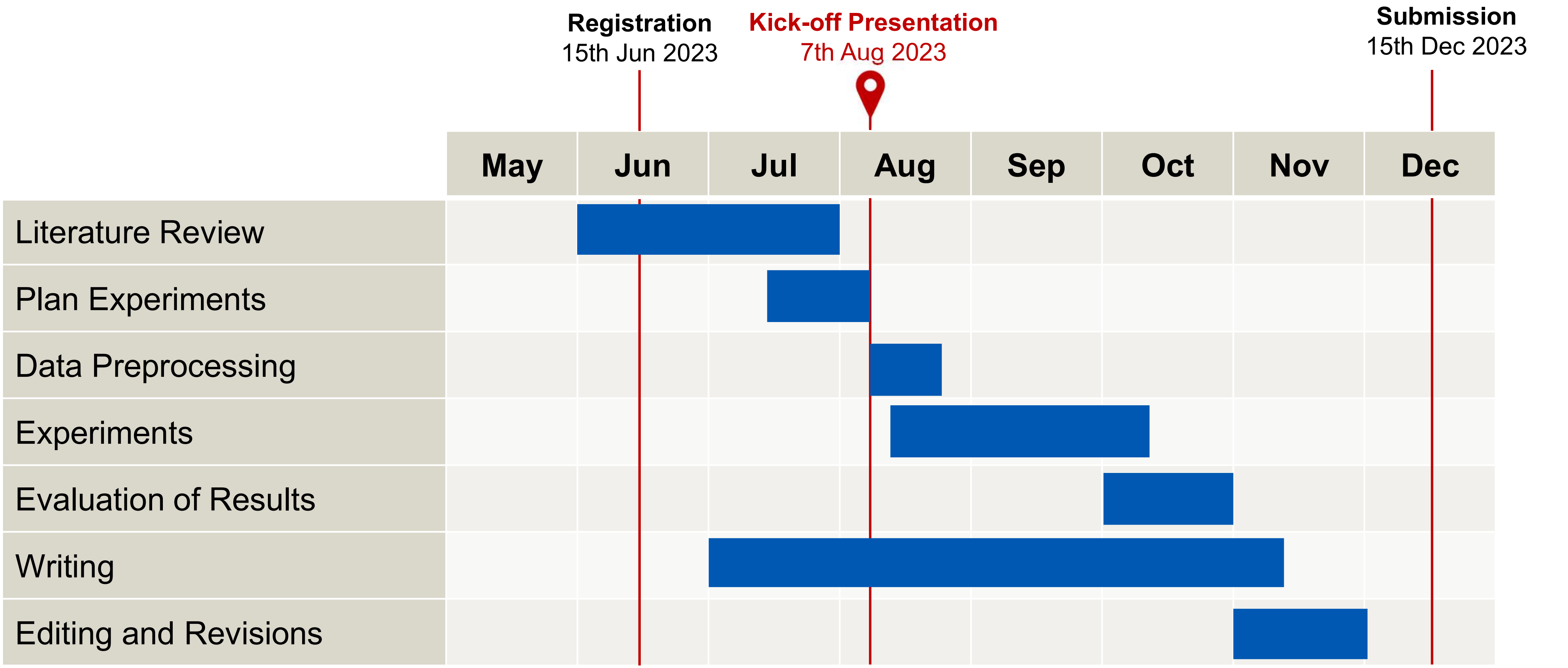
Document level privacy

Given any two input texts t, t', their corresponding outputs of the differential privacy mechanism ϕ_{priv} are $\boxed{\boldsymbol{\varepsilon * n}}$**-indistinguishable**.

Outline

1. Motivation
2. Research Questions
3. Methodology
4. Initial Findings
5. Next Steps
6. Timeline

Timeline


Registration
15th Jun 2023
Kick-off Presentation
7th Aug 2023
Submission
15th Dec 2023
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Literature Review
Plan Experiments
Data Preprocessing
Experiments
Evaluation of Results
Writing
Editing and Revisions

Alisha Riecker

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

alisha.riecker@tum.de
wwwmatthes.in.tum.de

Motivation - Differential Privacy for Word Embedding Vectors

$\boldsymbol{\Phi(w)}$

(Embedding Function)

+

$\boldsymbol{\eta}$

(Calibrated Random Noise)

=

$\boldsymbol{\Phi_{priv}(w)}$

(Randomized Mechanism)

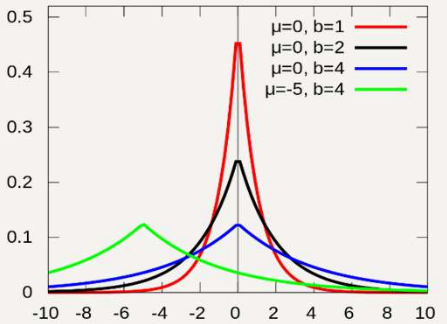
Example: Laplace Mechanism

An embedding function $\boldsymbol{\Phi(w)}$ with sensitivity Δ_ϕ :

$$\Delta_\phi = \max_{x,x' \epsilon X} \|\Phi(w) - \Phi(w')\|_1$$

Sensitivity describes the ***maximum possible change*** *in a function's output, when the input is changed.*

$\boldsymbol{\eta \sim Lap(\mu, b)}$,
where $\boldsymbol{\mu = 0}$, $\mathbf{b} = \boldsymbol{\frac{\Delta_\phi}{\varepsilon}}$



https://en.wikipedia.org/wiki/Laplace_distribution

Variance depends on sensitivity Δ_ϕ and privacy budget ε

Φ_{priv} is **$\boldsymbol{\varepsilon}$-differentially private** if for all pairs of inputs w, w' and all possible outputs z:

$$\frac{P[\Phi_{priv}(w) = z]}{P[\Phi_{priv}(w') = z]} \leq \exp(\varepsilon)$$

The probability of a specific output on any two inputs w, w' can differ by at most a ***multiplicative factor of $\boldsymbol{exp(\varepsilon)}$***.

Approaches for Bounding Sensitivity

$\Phi(w)$ has unbounded sensitivity → $\tilde{\Phi}(w) = f_{bound} \circ \Phi(w)$, s.th. $\tilde{\Phi}(w) \in [a, b]^d$

Approaches to bounding sensitivity		**Function f_{bound}**	**Sensitivity $\Delta_{\tilde{\Phi}}$**
Bound the embedding space	By observed range	*Define v_{max} as the maximum observed value.* $f_{bound}(y)_i = \begin{cases} y_i, & y_i \in [-v_{max}, v_{max}] \\ -v_{max}, & y_i < -v_{max} \\ v_{max}, & y_i > v_{max} \end{cases}$	$\Delta_{\tilde{\Phi}} \leq d * 2 * v_{max}$
	Normalize to $[0, 1]^d$	$f_{bound}(y) = \frac{y - \min(y)}{\max(y) - \min(y)}$	$\Delta_{\tilde{\Phi}} \leq d$
	Normalize to unit length	$f_{bound}(y) = \frac{y}{\lVert y \rVert}$	$\Delta_{\tilde{\Phi}} \leq 2$
Dimensionality Reduction	Johnson-Lindenstrauss Lemma	$f_{bound}(y) = A * x$, where $A \in \mathbb{R}^{k \times m}$ with a_{ij} drawn from $\left\{-\frac{1}{\sqrt{k}}, +\frac{1}{\sqrt{k}}\right\}$ iid. and $k = \left\lceil \frac{8 \ln(n)}{\varepsilon^2 - 2\frac{\varepsilon^3}{3}} \right\rceil$	$\Delta_{\tilde{\Phi}} \leq d' * (b - a)$, where $d' < d$

Approaches for Mapping to “real“ words

$\Phi_{priv}(w)$ does not correspond to a word embedding vector → $\Phi_{\boldsymbol{priv\_map}}(w) = f_{map} \circ \Phi_{priv}(w)$, s.th.,$\Phi_{\boldsymbol{priv\_map}}(w) \in Im(\Phi)$

Mapping	Function $\boldsymbol{f_{map}}$	Embedding Vector
No mapping	$f_{map}(y) = \mathrm{y}$	Noisy embedding vector
Map to nearest neighbor embedding	$f_{map}(y) = \underset{w \in \mathcal{W} \setminus \{x\}}{\operatorname{argmin}} \Vert y - \Phi(w) \Vert$	Similar embedding vector associated with "real" word
Randomly output first or second nearest neighbor	$f_{map}(y) = \begin{cases} w_1, & with\ prob.\, p \\ w_2, & with\ prob.\, 1 - p \end{cases}$ where $w_1 = \underset{w \in \mathcal{W} \setminus \{x\}}{\operatorname{argmin}} \Vert y - \Phi(w) \Vert$ $w_2 = \underset{w \in \mathcal{W} \setminus \{x, w_1\}}{\operatorname{argmin}} \Vert y - \Phi(w) \Vert$	Similar embedding vector associated with "real" word

Approaches for Bounding Sensitivity

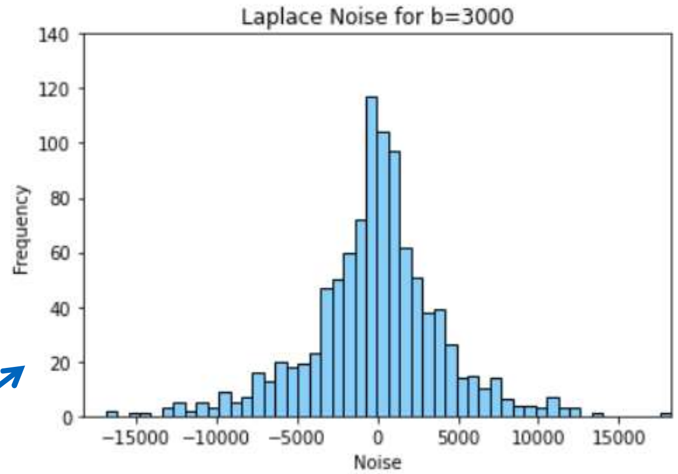
$\Phi(w)$ has unbounded sensitivity ➡ $\widetilde{\Phi}(w) = f_{bound} \circ \Phi(w),$ s.th. $\widetilde{\Phi}(w) \in [a,b]^d$

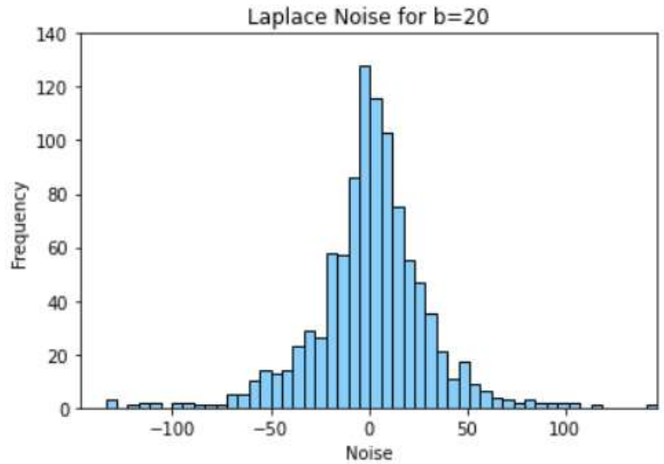
Privacy budget: $\varepsilon = 0.1$

Embedding function: $\Phi(w)\colon \mathcal{W} \to \mathbb{R}^{300}$

Noise: $\eta \sim Lap(\mu, b)$, where $\mu = 0$, $\mathrm{b} = \frac{\Delta_\Phi}{\varepsilon}$

Approaches to bounding sensitivity	Sensitivity $\Delta_{\widetilde{\Phi}}$	Variance of Laplace noise
Normalize to $[0,1]^d$	$\Delta_{\widetilde{\Phi}} \leq d$	$\frac{\Delta_\Phi}{\varepsilon} = 3000$
Normalize to unit length	$\Delta_{\widetilde{\Phi}} \leq 2$	$\frac{\Delta_\Phi}{\varepsilon} = 20$





Johnson-Lindenstrauss Lemma

Goal: Bound the worst-case distortion of distances during dimensionality reduction

By the Johnson-Lindenstrauss Lemma:

- $P = \{x_1, x_2, \dots, x_n\}$ a set of n points in $\mathbb{R}^m$
- $\xi \in (0, 1)$
- $k = \left\lceil \frac{8 \ln(n)}{\xi^2 - 2\frac{\xi^3}{3}} \right\rceil$

There exists a randomly generated mapping $f\colon \mathbb{R}^m \to \mathbb{R}^k$ such that **with probability at least $\mathbf{1 - \frac{1}{n^2}}$** :

$$\boldsymbol{(1-\xi)\left\|x_i - x_j\right\|_2 \leq \left\|f(x_i) - f(x_j)\right\|_2 \leq (1+\xi)\left\|x_i - x_j\right\|_2}$$

for all $i, j = 1, 2, \dots, n$.

Possible choice for mapping f:

$f(\mathrm{x}) = \mathrm{Ax}$, where A is a $k\ x\ m$ matrix with a_{ij} iid. from $\left\{-\frac{1}{\sqrt{k}}, +\frac{1}{\sqrt{k}}\right\}$

Unbounded Sensitivity

$\boldsymbol{\Phi(w)}$	+	$\boldsymbol{\eta}$	=	$\boldsymbol{\Phi_{priv}(w)}$
(Embedding Function)		(Calibrated Random Noise)		(Randomized Mechanism)

Sensitivity: $\Delta_\phi = \max_{w,w' \epsilon \mathcal{W}} \|\Phi(w) - \Phi\ (w')\|$

Problem: Unbounded Sensitivity

$\Phi(w)\ \epsilon\ (-\infty, +\infty)^d \quad \Rightarrow \quad \boldsymbol{\Delta_\phi} = \infty$

Approach: Bound Sensitivity

$\widetilde{\Phi}(w) = f_{bound} \circ \Phi(w),$ s.th. $\widetilde{\Phi}(w) \in\ [a, b]^d \quad \Rightarrow \quad \boldsymbol{\Delta_{\widetilde{\Phi}} = d * (b - a)}\ < \infty$

$\boldsymbol{\Phi}_{priv}(w)$ is not a “real” word

$\boldsymbol{\Phi}(\boldsymbol{w})$	+	$\boldsymbol{\eta}$	=	$\boldsymbol{\Phi}_{\boldsymbol{priv}}(\boldsymbol{w})$
(Embedding Function)		(Calibrated Random Noise)		(Randomized Mechanism)

Problem: $\boldsymbol{\Phi}_{\boldsymbol{priv}}(\boldsymbol{w})$ is not a “real” word

$\Phi_{\boldsymbol{priv\_map}}(w) \notin Im(\Phi)$

Approach: Map $\boldsymbol{\Phi}_{\boldsymbol{priv}}(\boldsymbol{w})$ to similar embedding vectors associated with real words

$\Phi_{\boldsymbol{priv\_map}}(w) = f_{map} \circ \Phi_{priv}(w)$, s.th.,$\Phi_{\boldsymbol{priv\_map}}(w) \in Im(\Phi)$