Author Obfuscation using Generalised Differential Privacy

Natasha Fernandes, Mark Dras, Annabelle McIver

Macquarie University, Sydney, Australia

Abstract. The problem of obfuscating the authorship of a text document has received little attention in the literature to date. Current approaches are ad-hoc and rely on assumptions about an adversary's auxiliary knowledge which makes it difficult to reason about the privacy properties of these methods. Differential privacy is a well-known and robust privacy approach, but its reliance on the notion of adjacency between datasets has prevented its application to text document privacy. However, generalised differential privacy permits the application of differential privacy to arbitrary datasets endowed with a metric and has been demonstrated on problems involving the release of individual data points. In this paper we show how to apply generalised differential privacy to author obfuscation by utilising existing tools and methods from the stylometry and natural language processing literature.

Keywords: generalised differential privacy, author obfuscation, word mover's distance

1 Introduction

The proliferation of machine learning techniques and publicly available datasets has resulted in some exciting advances in data analytics. At the same time, some well-publicised privacy breaches, notably the AOL and Netflix examples, have created concerns about data privacy and the ability for privacy methods to protect against machine learning attacks. Differential privacy is a key privacy definition which has rapidly gained popularity due to its mathematical foundations and, importantly, its independence from assumptions about the external data sources available to adversaries. This makes it an important tool for protection of personal data in the face of machine learning adversaries trained on large-scale datasets.

Differential privacy relies on a notion of an 'individual' in a dataset, under the assumption that datasets are structured into rows of individuals, and adversaries are agents which query the data for statistical information. These notions are not naturally transferable to unstructured datasets such as text documents.

Generalised differential privacy is an extension of differential privacy which can be applied to arbitrary datasets endowed with a metric. This permits its application to more general datasets, and has found most application in geo-location privacy, involving the differentially private release of users'

We recall the usual definition of differential privacy which says that, for any pair of adjacent datasets x, x' and some output z from a query, a (probabilistic) mechanism K satisfies ϵ-differential privacy if

$$K(x)(z) \leq e^\epsilon K(x')(z)$$

for some non-negative ϵ. [1]

Generalised differential privacy extends this notion to domains endowed with a metric as follows: for any domain of secrets $\mathcal{X}$ endowed with a metric $d_\mathcal{X}$ and any elements $x, x' \in \mathcal{X}$, a (probabilistic) mechanism K satisfies $\epsilon d_\mathcal{X}$-privacy if

$$K(x)(z) \leq e^{\epsilon d_\mathcal{X}(x,x')} K(x')(z)$$

In this paper we show how generalised differential privacy can be applied to author obfuscation. This task requires the private release of documents so as to protect the identity of the author whilst maintaining some semantic properties of the document. We draw on existing notions of authorship from the stylometry and natural language processing literature which incorporate the use of distance measures between authors.

2 Problem Description

Author obfuscation is the task of obscuring a piece of text in order to hide its authorship whilst preserving its semantic content. Stylometric techniques have identified three types of features used to identify authorship of a document: stylistic, word-based and character-based features. Stylistic features are typically lexical, syntactic or document-level characteristics. For example, average word length, average sentence length and frequency of use of particular words are all features which can be unique for authors. Word-based methods treat each word in the document as a feature, and represent a document as a *bag of words*, which ignores word ordering but preserves frequency counts of individual words. Finally, character-based features treat individual sequences of characters as features for document representation. These sequences are referred to as *character n-grams*. For example, the character 3-gram representation of the phrase "There it is" would be 'The', 'her', 'ere', 're_', 'e_it', 'it_', 't_i', '_is' [2].

2.1 Model

We envisage an author who wishes to release a document which preserves the topicality of the original document whilst masking stylometric features which may reveal their authorship. By 'topicality' we refer to the ability of a document consumer to identify the original topic of the document. In order to provide a

---

[1] We use only the strict $\{\epsilon, 0\}$ version of differential privacy in this paper, and not its relaxation (ϵ, δ)-differential privacy.

[2] Note that we use '_' to represent spaces.

privacy guarantee for any adversary (ie over any prior), we will use generalised differential privacy. Our privacy promise is that the output document is almost the same, regardless of whether the input document was x or some 'close' document x', where closeness is defined using an appropriate metric distance. In order to achieve privacy over authors, we need to modify the features in the document which reveal authorship. This can be done firstly by considering documents formatted as bags of words (BOW). Such document are useful for machine learning applications, which typically use BOW formats and ignore word ordering. This formatting also removes word ordering information, which reveals important stylistic information. Secondly, we can remove 'stopwords', which are words such as pronouns and prepositions, which do not contribute to the meaning of the document. These have been shown to be highly effective in author attribution, but because they contribute nothing to our utility requirement they can be safely discarded. Finally, we can consider the removal of words which do not significantly contribute to the topicality of the document, so as to reduce the document down to the smallest set of words which guarantee some usefulness. This can be done using a feature classifier to identifier the most significant features for topic classification.

2.2 Application of Differential Privacy

We can represent words as real-valued vectors using a word embedding representation such as Word2Vec or GloVe. These representations preserve semantic relationships between words, where the semantic distance is typically measured using either cosine similarity or Euclidean distance. This also allows the entire word embedding vocabulary to also be treated as a synonym set for any word in the vocabulary. A natural metric to then consider for measuring the semantic distance between documents is the Word Mover's Distance. This metric has been designed specifically for use with word embedding vectors, and is based on the well-known Earth Mover's Distance. Briefly, the Word Mover's Distance is the cost of moving all the words from one document to the words in another document. We will formalise this in the next section.

3 Preliminaries

In this section we lay out some definitions for use throughout the rest of this paper. We will only be interested in discrete sets so we present a simplified formalisation, noting that the definition also applied to continuous sets and distributions.

Let $\mathcal{X}$ and $\mathcal{Z}$ be finite sets and let $\mathbb{P}(\mathcal{Z})$ be the set of probability measures over $\mathcal{Z}$. We define a *mechanism* as a probabilistic function $K : \mathcal{X} \to \mathbb{P}(\mathcal{Z})$

Recall that a metric $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ satisfies (i) $d(x, y) = 0$ iff $x = y$, (ii) $d(x, y) = d(y, x)$ and (iii) $d(x, y) + d(y, z) \geq d(x, z)$ for all $x, y, z \in \mathcal{X}$. We denote by d_2 the Euclidean metric on $\mathbb{R}^n$.

We recall the definition of generalised differential privacy:

Definition 1. *(Generalised Differential Privacy) Let $\epsilon > 0$. A mechanism K : $\mathcal{X} \to \mathbb{P}(\mathcal{Z})$ satisfies $d_{\mathcal{X}}$-privacy, iff $\forall x, x' \in \mathcal{X}$:*

$$K(x)(Z) \le e^{\epsilon d_{\mathcal{X}}(x,x')} K(x')(Z) \quad \forall Z \subseteq \mathcal{Z} \text{ }^3$$

We now formalise some notions from the natural language processing literature.

Let $\mathcal{V}$ be a fixed finite vocabulary of words from all possible documents. A *bag of words* (BOW) is an unordered, finite-length lists of words from $\mathcal{V}$ with duplicates permitted. A *document vector* is an ordered bag of words. A *word embedding vector* is a k-dimensional real-valued vector representing a word in $\mathcal{V}$, for some fixed positive integer k. We assume the existence of a word embedding vector lookup table, denoted W, such that $W(w)$ returns the word embedding vector for the word $w \in \mathcal{V}$.

We denote by $\mathcal{W}$ the universe of word embedding vectors (for all words in $\mathcal{V}$). We denote by $\mathcal{U}(\mathcal{W})$ the set of unordered lists of word vectors, also known as BOWs, and by $\mathcal{O}(\mathcal{W})$ the set of ordered lists of word vectors, also known as document vectors.

We note that we can transform a BOW into a document vector by fixing an (arbitrary) ordering of words.

The Word Mover's Distance can be formally defined as follows:

Definition 2. *(Word Mover's Distance) Let $x, y \in \mathcal{O}(\mathcal{W})$ be document vectors of lengths a and b respectively. We assume the existence of a non-negative, real-valued cost function over $\mathcal{V}$. Let $C \in \mathbb{R}^{a \times b}$ be a cost matrix, where C_{ij} represents the cost of moving word i in x to word j in y. Define $T \in \mathbb{R}^{a \times b}$ to be a flow matrix where the entry T_{ij} denotes how much of word i in x moves to word j in y. Then the Word Mover's Distance $d_W(x, y)$ is defined as the solution to the linear optimisation problem:*

$$d_W(x, y) = \min_{T \ge 0} \sum_{i,j} T_{ij} C_{ij}$$

subject to:

$$\sum_j T_{ij} = \frac{1}{a} \qquad \forall i \in [1 \ldots a]$$

and

$$\sum_i T_{ij} = \frac{1}{b} \qquad \forall j \in [1 \ldots b]$$

We also define our notion of document privacy under the term *document-indistinguishability*.

Definition 3. *(Document-Indistinguishability) Let $d_W : \mathcal{U}(\mathcal{W}) \times \mathcal{U}(\mathcal{W}) \to [0, \infty)$ be the Word Mover's Distance defined on BOW documents. A mechanism K : $\mathcal{U}(\mathcal{W}) \to \mathbb{P}(\mathcal{U}(\mathcal{W}))$ satisfies ϵ-document-indistinguishability iff for all $x, x' \in \mathcal{U}(\mathcal{W})$:*

$$K(x)(Z) \le e^{\epsilon d_W(x,x')} K(x')(Z) \quad \forall Z \subseteq \mathcal{U}(\mathcal{W})$$

---

[3] We make the simplifying assumption that all sets in $\mathcal{Z}$ are measurable.

4 Privacy Using the Word Mover's Distance

We now present some results on the Word Mover's Distance which will allow us to produce a differentially private mechanism for documents.

4.1 Optimal Solution to the Word Mover's Distance

Our first result shows that an optimal solution to the Word Mover's Distance problem involves the movement of whole words only, precisely when the source and destination documents are the same length. In order to prove this result, we first introduce some results on doubly stochastic matrices.

Definition 4. *An $n \times n$ matrix whose elements are non-negative and has all rows and columns summing to 1 is called* doubly stochastic. *A doubly stochastic matrix which contains only 1's and 0's is called a* permutation matrix.

Theorem 1. *(Birkhoff-von Neumann) The set of $n \times n$ doubly stochastic matrices forms a convex polytope whose vertices are the $n \times n$ permutation matrices.*

The Birkhoff-von Neumann theorem says that the set of doubly stochastic matrices is a closed, bounded convex set, and every doubly stochastic matrix can be written as a convex combination of the permutation matrices. We can use this theorem to prove the following result.

Theorem 2. *Let C be an $n \times n$ cost matrix. Then the optimisation problem minimise $\sum_{i,j} T_{ij} C_{ij}$ where T is an $n \times n$ doubly stochastic matrix always has an $n \times n$ permutation matrix as an optimal solution.*

We are now ready to present a result on the Word Mover's Distance, namely that documents of equal length always have an optimal solution which does not involve partial movements of words. We note that the order of words in a document does not affect the calculation of the Word Mover's Distance, hence the following result on document *vectors* also holds for BOW documents.

Theorem 3. *Let $d_1, d_2 \in \mathcal{O}(\mathcal{W})$ be n-dimensional document vectors. Then the Word Mover's Distance $d_W(d_1, d_2)$ has an optimal solution involving only the movement of whole words.*

Proof. Since d_1 and d_2 are both of length n, the flow matrix T must be $n \times n$ so we can rewrite the constraints as

$$\sum_i T_{ij} = \frac{1}{n} \quad \forall j \in [1 \ldots n] \quad \text{and} \quad \sum_j T_{ij} = \frac{1}{n} \quad \forall i \in [1 \ldots n]$$

Notice that the optimisation problem is in essence unchanged if we multiply T by a constant, so we could also write

$$\sum_i T_{ij} = 1 \quad \forall j \in [1 \ldots n] \quad \text{and} \quad \sum_j T_{ij} = 1 \quad \forall i \in [1 \ldots n]$$

Thus we have a T that is doubly stochastic. From Theorem 2 the optimal solution includes a permutation matrix, that is, a matrix in which there is exactly one 1 in each row and column and the remaining elements are 0. But this corresponds to a flow where each word in d_1 moves entirely to a whole word in d_2. This completes the proof. □

Example We will use a simple example to demonstrate how the flow matrix can be simplified when both documents are of the same length. Consider the documents $\boldsymbol{d}$ = 'Obama speaks Illinois' and $\boldsymbol{d'}$ = 'President greets press'. The relative mass of each word in each document is $\frac{1}{3}$ since both documents are the same length; this is depicted by the vectors $\boldsymbol{d_p}$ and $\boldsymbol{d'_p}$ in Figure 3. We imagine a cost matrix such that the flow matrix T given in Figure 1 is optimal. Now, consider the scenario in Figure 2 where the relative mass of each word is 1. Since we have simply multiplied the relative weights of all words by 3 without changing the cost matrix, the optimal solution matrix T' will correspond exactly to $3T$, and becomes doubly stochastic. Although the computed distance will also be 3 times the original distance, the *relative* flow between words in the 2 documents is preserved. In particular, whole word flows in Figure 1 correspond with whole word flows in Figure 2. Thus the result of Theorem 3 stands in both cases.
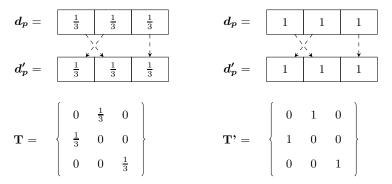
$$\boldsymbol{d_p} = \begin{array}{|c|c|c|} \hline \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline \end{array} \qquad \boldsymbol{d_p} = \begin{array}{|c|c|c|} \hline 1 & 1 & 1 \\ \hline \end{array}$$

$$\boldsymbol{d'_p} = \begin{array}{|c|c|c|} \hline \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline \end{array} \qquad \boldsymbol{d'_p} = \begin{array}{|c|c|c|} \hline 1 & 1 & 1 \\ \hline \end{array}$$

$$\mathbf{T} = \left\{ \begin{array}{ccc} 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{3} \end{array} \right\} \qquad \mathbf{T'} = \left\{ \begin{array}{ccc} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{array} \right\}$$

Fig. 1: Sample vectors and flow matrix T for standard Word Mover's Distance formulation.

Fig. 2: Modified vectors using multiplicative factor of 3 to produce doubly stochastic flow matrix T'.

Fig. 3: Example of modifying the Word Mover's Distance problem to generate a doubly stochastic flow matrix. Multiplying the relative word weights by a constant corresponds to multiplication of the flow matrix by the same constant, however the relative flows between words remains unchanged.

4.2   Extension to Document-Indistinguishability

We are now ready to present our main theorem, which provides a connection between generalised differential privacy for vectors and document indistinguishability.

Theorem 4. *Let $K : \mathbb{R}^k \rightarrow \mathbb{P}(\mathbb{R}^k)$ be a mechanism operating on real-valued k-dimensional vectors which satisfies ϵd_2-privacy. Then K can be extended to a mechanism $K^* : \mathcal{U}(\mathcal{W}) \rightarrow \mathbb{P}(\mathcal{U}(\mathcal{W}))$ operating on documents which satisfies ϵ-document-indistinguishability.*

Proof Sketch Firstly, we can think of K as a mechanism operating on document (ordered) vectors, and show that it can be extended to a mechanism operating on BOW documents (unordered vectors) by considering the appropriate permutations. It turns out that we can fix any ordering of words in the source document and just consider permutations of the output document. This means the privacy guarantee for input documents x, x' is determined by any ordering we choose. We can therefore choose an ordering which minimises the transportation cost between corresponding words in the documents, which, from Theorem 3, corresponds to the WMD.

5   Privacy Mechanism

We are now ready to present a privacy mechanism for document-indistinguishability. We have seen that to find a mechanism satisfying ϵ-document-indistinguishability, it suffices to find a mechanism operating on vectors which satisfies ϵd_2-privacy. Previous work [1] has shown how this can be done in 2-dimensions via the planar Laplacian. We now present an extension of this result to n-dimensions.

5.1   n-Dimensional Laplace Mechanism

Given sets $\mathcal{X}, \mathcal{Z}$ of n-dimensional vectors, we would like a mechanism K with pdf D satisfying

$$D(x)(z) \propto e^{-\epsilon d_2(x,z)} \quad \text{for } x \in \mathcal{X}, z \in \mathcal{Z}$$

Such a mechanism is called a Laplace mechanism and satisfies ϵd_2-privacy [2]. We require a method of selecting a vector according to this distribution.

Noting that D is spherically symmetric, and using translation invariance, we can consider the distribution $D(0)(z)$ and translate this by x to get the distribution $D(x)(z)$. For notational convenience we write $D(0)(z)$ as $D_0(z)$.

Using $z = (z_1, z_2, \ldots, z_n)$ then $D_0(z)$ can be rewritten as

$$D_0(z) = ce^{-\epsilon\sqrt{z_1^2 + z_2^2 + \ldots + z_n^2}}$$

Calculating the constant c can be done by a mapping $(z_1, z_2, \ldots, z_n) \mapsto (r, \theta_1, \theta_2, \ldots, \theta_{n-1})$ to spherical co-ordinates. This yields the following integral

$$\int_0^\infty c_1 r^{n-1} e^{-\epsilon r} dr \int_0^\pi c_2 \sin^{n-2} \theta_1 d\theta_1 \int_0^\pi c_3 \sin^{n-3} \theta_2 d\theta_2 \cdots \int_0^\pi c_{n-1} \sin \theta_{n-2} d\theta_{n-2} \int_0^{2\pi} c_n d\theta_{n-1}$$

where $\prod_{k=1}^n c_k = c$.

This is a product of independent distributions, and in particular, the constant c_1 in the first integral can be shown to evaluate to $\frac{\epsilon^n}{(n-1)!}$ yielding

$$\int_0^\infty \frac{\epsilon^n}{(n-1)!} r^{n-1} e^{-\epsilon r} dr = \int_0^\infty \frac{\epsilon^n}{\Gamma(n)} r^{n-1} e^{-\epsilon r} dr$$

which we recognise as the PDF of the Gamma distribution

$$f(x, k; \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

where $x \mapsto r$, $k \mapsto n$ and $\theta \mapsto \frac{1}{\epsilon}$.

The remaining product of integrals is equivalent to selecting an n-dimensional vector uniformly over the unit n-sphere. Therefore, the selection of a random n-dimensional Laplace vector can be achieved by selecting a random vector uniformly over the surface of an n-sphere and applying a scaling factor drawn from the gamma distribution. We formalise this as follows

Theorem 5. *Let U_{S_1} be the uniform distribution on the n-dimensional sphere of radius 1, and denote by $Gamma(k, \theta)$ the Gamma distribution with shape k and scale θ. Let $K : \mathbb{R}^n \to \mathbb{P}(\mathbb{R}^n)$ be a mechanism operating on real-valued n-dimensional vectors which outputs z with distribution $x + UR$, where $U \sim U_{S_1}$ and $R \sim Gamma(n, \frac{1}{\epsilon})$. Then the mechanism K is a Laplace mechanism satisfying ϵd_2-privacy.*

Note that choosing n=2 results in the planar Laplacian described in [1].

Several methods have been proposed for the generation of random variables from the Gamma distribution [3] as well as the uniform selection of points on the unit n-sphere [4]. We will not present methods for the former, as there are already implementations in libraries such as Scipy (for Python) for selecting from the Gamma distribution. However, there is a nice method for selecting a random vector from the surface of the unit n-sphere which has been described previously in the literature [4]. The method is to select n random variables from the Gaussian distribution over $[0, 1]$ into an n-dimensional vector v and output the normalised vector $\frac{v}{|v|}$. This method allows a random unit n-vector to be drawn without requiring a transformation from polar co-ordinates.

5.2 Mechanism for Document Indistinguishability

We now present a mechanism satisfying ϵ-indistinguishability. We assume that input documents have first been converted into a bag of words with stopwords discarded. We then use the method for generating fixed-length documents described in [6]. That is, the bag of words document can be used to generate a distribution over words using the frequency count of each word in the document. A fixed-length document can be generating by selecting n words from the document according to the distribution. These preprocessing steps d

This is shown in Algorithm 1.

Algorithm 1 Obfuscation Mechanism

Require: epsilon ϵ, word embeddings W, documents d
  for doc in d **do**
    words = list words in doc
    for w in words **do**
      x = lookup vector for w in W
      r = select scale according to Gamma(dim(x), $\frac{1}{\epsilon}$)
      u = select unit vector uniformly on dim(x)-sphere
      z = x + ru
      z' = lookup closest word to z in W
      add z' to noisy_doc
    end for
    add noisy_doc to obfuscated dataset
  end for
  return obfuscated dataset

Theorem 6. *The mechanism presented in Algorithm 1 satisfies ϵ-document-indistinguishability.*

Proof. The inner loop contains the Laplace mechanism as described in Theorem 5, which satisfies ϵd_2-privacy. The 'closest word' step represents a postprocessing truncation of the vector z, which does not change the ϵd_2-privacy guarantee of the inner loop. The outer loop applies mechanism in the inner loop to every word in the document. By Theorem 4 this outer loop satisfies ϵ-document-indistinguishability. $\qquad\qquad\square$

A   Proofs Omitted from Section 4

Theorem 7. *Let C be an $n \times n$ cost matrix. Then the optimisation problem minimise $\sum\limits_{i,j} T_{ij} C_{ij}$ where T is an $n \times n$ doubly stochastic matrix always has an $n \times n$ permutation matrix as an optimal solution.*

Proof. We prove this by contradiction. Let T^* be an optimal $n \times n$ solution matrix. We know that such a solution exists by the Birkhoff-von Neumann Theorem (since the set of solutions is closed and bounded). We assume firstly that T^* is not a permutation matrix, and secondly that no permutation matrix is optimal. Let $\{P^1, P^2, \ldots, P^k\}$ be the set of $n \times n$ permutation matrices. Then, by the Birkhoff-von Neumann theorem, we can write

$$T^* = \lambda_1 P^1 + \lambda_2 P^2 + \ldots + \lambda_k P^k \tag{1}$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^{k} \lambda_i = 1$. Since T^* is optimal and none of the P^i are optimal, we can also write

$$\sum_{i,j} P_{ij}^m C_{ij} > \sum_{i,j} T_{ij}^* C_{ij} \quad \text{(by assumption)}$$

for $0 < m \leq k$. And thus we have

$$\sum_{i,j} T_{ij}^* C_{ij} = \sum_{i,j} (\lambda_1 P_{ij}^1 + \ldots + \lambda_k P_{ij}^k) C_{ij} \quad \text{(from 1)}$$

$$= \sum_{i,j} \lambda_1 P_{ij}^1 C_{ij} + \ldots + \sum_{i,j} \lambda_k P_{ij}^k C_{ij} \quad \text{(factorising)}$$

$$> \sum_{i,j} \lambda_1 T_{ij}^* C_{ij} + \ldots + \sum_{i,j} \lambda_k T_{ij}^* C_{ij} \quad \text{(by assumption)}$$

$$= \lambda_1 \sum_{i,j} T_{ij}^* C_{ij} + \ldots + \lambda_k \sum_{i,j} T_{ij}^* C_{ij} \quad \text{(arithmetic)}$$

$$= (\lambda_1 + \ldots + \lambda_k) \sum_{i,j} T_{ij}^* C_{ij} \quad \text{(factorising)}$$

$$= \sum_{i,j} T_{ij}^* C_{ij} \quad \left(\text{since } \sum_{i=1}^{k} \lambda_i = 1\right)$$

which is a contradiction. Thus, either T^* is a permutation matrix, or there must be a permutation matrix which is also optimal.

The following lemma is useful in proving the next main result on document indistinguishability.

Lemma 1. *Let $K : \mathcal{W} \to \mathbb{P}(\mathcal{W})$ be a mechanism operating on word vectors and let $d, z \in \mathcal{U}(\mathcal{W})$ be documents of length n. Then K can be extended to a mechanism $K^* : \mathcal{U}(\mathcal{W}) \to \mathbb{P}(\mathcal{U}(\mathcal{W}))$ operating on documents such that*

$$K^*(d)(z) = \sum_i K(w_1)(v_{\phi_i(1)}) \times K(w_2)(v_{\phi_i(2)}) \times \ldots \times K(w_n)(v_{\phi_i(n)})$$

where the w_i, v_i are words (arbitrarily labelled) in d, z respectively, the ϕ_i are permutation functions, and the sum is over unique permutations of words in z.

Proof. Choose an arbitrary ordering of words in d and z and let $\boldsymbol{d} =< w_1, w_2, \ldots, w_n >$, $\boldsymbol{z} =< v_1, v_2, \ldots, v_n >$ be the corresponding document vectors. Let $\phi_i : S \to S$ be a sequence of permutation functions over $S = \{1, \ldots, n\}$ for $i = \{1, \ldots, m\}$ such that $< v_{\phi_i(1)}, v_{\phi_i(2)}, \ldots, v_{\phi_i(n)} >$ is a unique permutation of words in z for each i.

Now, it is straightforward to extend K to a mechanism $K' : \mathcal{O}(\mathcal{W}) \to \mathbb{P}(\mathcal{O}(\mathcal{W}))$ operating on document *vectors*, since we can simply apply K to each word in order. That is,

$$K'(\boldsymbol{d})(\boldsymbol{z}) = K(w_1)(v_1) \times K(w_2)(v_2) \times \ldots \times K(w_n)(v_n)$$

Clearly, $K'(\boldsymbol{d})$ defines a valid probability distribution for any $\boldsymbol{d}$ since we sum over all possible output vectors $\boldsymbol{z}$.

In order to extend this to a mechanism over documents, observe that the mechanism K' produces the same output distribution regardless of the ordering of words in the document vector $\boldsymbol{d}$ (since the mechanism K operates on each word independently). Therefore we only need to consider permutations of words in the output document vector $\boldsymbol{z}$. The distribution over documents is then given by the sum of distributions over each permutation of words in the output vector $\boldsymbol{z}$, that is,

$$K^*(d)(z) = \sum_i K(w_1)(v_{\phi_i(1)}) \times K(w_2)(v_{\phi_i(2)}) \times \ldots \times K(w_n)(v_{\phi_i(n)})$$

Clearly $K^*(d)$ also defines a valid probability distribution, since it produces the same distribution as $K'(\boldsymbol{d})$ except that the output probabilities are 'collected' for all permutations of the output vector. $\square$

Theorem 8. *Let $K : \mathbb{R}^k \to \mathbb{P}(\mathbb{R}^k)$ be a mechanism operating on real-valued k-dimensional vectors which satisfies ϵd_2-privacy. Then K can be extended to a mechanism $K^* : \mathcal{U}(\mathcal{W}) \to \mathbb{P}(\mathcal{U}(\mathcal{W}))$ operating on documents which satisfies ϵ-document-indistinguishability.*

Proof. Since we have a real-valued vector representation for words, we can treat K as a mechanism operating on word vectors. Let $\boldsymbol{d} =< w_1, w_2, \ldots, w_n >$, $\boldsymbol{z} =< v_1, v_2, \ldots, v_n >$ be n-dimensional document vectors and let $\phi_i : S \to S$ be a sequence of permutation functions over $S = \{1, \ldots, n\}$ for $i = \{1, \ldots, m\}$. From Lemma 1, we can extend K to a mechanism $K^* : \mathcal{U}(\mathcal{W}) \to \mathbb{P}(\mathcal{U}(\mathcal{W}))$ satisfying

$$K^*(\boldsymbol{d})(\boldsymbol{z}) = \sum_i K(w_1)(v_{\phi_i(1)}) \times K(w_2)(v_{\phi_i(2)}) \times \ldots \times K(w_n)(v_{\phi_i(n)})$$

We can choose any particular ordering of words in $\boldsymbol{d}$ since the ordering of words is arbitrary. Fix any ordering of words in $\boldsymbol{d}$ and let $\boldsymbol{d}' =< w_1', w_2', \ldots, w_n' >$ be an n-dimensional document vector, where the word order in $\boldsymbol{d}'$ is chosen to minimise the sum of the Euclidean distances between corresponding words in $\boldsymbol{d}$ and $\boldsymbol{d}'$. That is, we choose an ordering which minimises $\sum\limits_{i=1}^{n} d_2(w_i, w_i')$.

Now, for the document vectors $\boldsymbol{d}$, $\boldsymbol{d'}$, we have that

$$\frac{K^*(\boldsymbol{d})(\boldsymbol{z})}{K^*(\boldsymbol{d'})(\boldsymbol{z})} = \frac{\sum_i K(w_1)(v_{\phi_i(1)}) \times K(w_2)(v_{\phi_i(2)}) \times \ldots \times K(w_n)(v_{\phi_i(n)})}{\sum_i K(w'_1)(v_{\phi_i(1)}) \times K(w'_2)(v_{\phi_i(2)}) \times \ldots \times K(w'_n)(v_{\phi_i(n)})}$$

$$= \frac{\sum_i \prod_j K(w_j)(v_{\phi_i(j)})}{\sum_i \prod_j K(w'_j)(v_{\phi_i(j)})} \tag{2}$$

where the number of terms in the numerator and denominator is the same (since this depends only on $\boldsymbol{z}$). But we also have that

$$K(w_k)(v_{\phi_i(k)}) \leq K(w'_k)(v_{\phi_i(k)})e^{\epsilon d_2(w_k, w'_k)} \quad (\epsilon d_2\text{-privacy}) \tag{3}$$

for all words in $\boldsymbol{d}$, $\boldsymbol{d'}$ and all permutations ϕ_i. Therefore,

$$\frac{K^*(\boldsymbol{d})(\boldsymbol{z})}{K^*(\boldsymbol{d'})(\boldsymbol{z})} = \frac{\sum_i \prod_j K(w_j)(v_{\phi_i(j)})}{\sum_i \prod_j K(w'_j)(v_{\phi_i(j)})} \quad (\text{from } 2)$$

$$\leq \frac{\sum_i \prod_j K(w'_j)(v_{\phi_i(j)})e^{\epsilon d_2(w_j, w'_j)}}{\sum_i \prod_j K(w'_j)(v_{\phi_i(j)})} \quad (\text{from } 3)$$

$$= \frac{\sum_i e^{\epsilon(d_2(w_1, w'_1) + \ldots + d_2(w_n, w'_n))} \prod_j K(w'_j)(v_{\phi_i(j)})}{\sum_i \prod_j K(w'_j)(v_{\phi_i(j)})} \quad (\text{arithmetic})$$

$$= \frac{e^{\epsilon(d_2(w_1, w'_1) + \ldots + d_2(w_n, w'_n))} \sum_i \prod_j K(w'_j)(v_{\phi_i(j)})}{\sum_i \prod_j K(w'_j)(v_{\phi_i(j)})} \quad (\text{arithmetic})$$

$$= e^{\epsilon(d_2(w_1, w'_1) + \ldots + d_2(w_n, w'_n))} \quad (\text{cancelling like terms}) \tag{4}$$

Now, notice that the documents $\boldsymbol{d}$ and $\boldsymbol{d'}$ have the same dimension (necessarily, due to the operation of the mechanism K^*). Therefore we know from Theorem 3 that the Word Mover's Distance $d_{\mathcal{W}}(\boldsymbol{d}, \boldsymbol{d'})$ has an optimal solution involving the movement of whole words. That is, there exists a permutation of words $< w'_{\phi_i(1)}, w'_{\phi_i(2)}, \ldots, w'_{\phi_i(n)} >$ in $\boldsymbol{d'}$ such that $d_{\mathcal{W}}(\boldsymbol{d}, \boldsymbol{d'}) = \sum_k d_2(w_k, w'_{\phi_i(k)})$. But we chose an ordering of words in $\boldsymbol{d'}$ that minimises $\sum_k d_2(w_k, w'_k)$. Recalling that the Word Mover's Distance is minimal, we therefore must have that

$$d_{\mathcal{W}}(\boldsymbol{d}, \boldsymbol{d'}) = \sum_k d_2(w_k, w'_k)$$

And so,

$$\frac{K^*(\boldsymbol{d})(\boldsymbol{z})}{K^*(\boldsymbol{d'})(\boldsymbol{z})} \leq e^{\epsilon(d_2(w_1, w'_1) + \ldots + d_2(w_n, w'_n))} \quad (\text{from } 4)$$

$$= e^{\epsilon d_W(\boldsymbol{d}, \boldsymbol{d'})}$$

Thus the mechanism K^* satisfies ϵ-document-indistinguishability. $\square$

B   Proofs Omitted from Section 5

We present here a more complete proof of the derivation of the n-dimensional Laplace mechanism.

We consider the distribution $D_0(z) = ce^{-\epsilon\sqrt{z_1^2+z_2^2+\cdots+z_n^2}}$ for some constant c. In order to select a point from this distribution, we consider the CDF

$$F(z) = \int \cdots \int_{Z_A} ce^{-\epsilon\sqrt{z_1^2+z_2^2+\cdots+z_n^2}}\, dz_1 \ldots dz_n \tag{5}$$

for some region of interest Z_A.

To compute this we require a change of co-ordinates. We can convert from Cartesian co-ordinates to spherical co-ordinates

$$(z_1, z_2, z_3, \ldots, z_n) \mapsto (r, \theta_1, \theta_2, \ldots, \theta_{n-1})$$

as stated in [5] using a transformation $r = \sqrt{z_1^2 + z_2^2 + \ldots + z_n^2}$ with inverse

$$z_1 = r\cos\theta_1$$
$$z_2 = r\sin\theta_1\cos\theta_2$$
$$z_3 = r\sin\theta_1\sin\theta_2\cos\theta_3$$
$$\ldots$$
$$z_{n-1} = r\sin\theta_1\sin\theta_2\ldots\sin\theta_{n-2}\cos\theta_{n-1}$$
$$z_n = r\sin\theta_1\sin\theta_2\ldots\sin\theta_{n-2}\sin\theta_{n-1}$$

We also need to calculate the matrix of partial derivatives to get the Jacobian, which is well known to be

$$\frac{\partial(z_1, z_2, \ldots, z_n)}{\partial(r, \theta_1, \ldots, \theta_{n-1})} = r^{n-1}\sin^{n-2}\theta_1\sin^{n-3}\theta_2\ldots\sin^2\theta_{n-3}\sin\theta_{n-2}$$

And therefore the integral in (5) becomes

$$\int \cdots \int_{Z_A} ce^{-\epsilon\sqrt{z_1^2+z_2^2+\cdots+z_n^2}}\, dz_1 \ldots dz_n$$

$$= \int \cdots \int_{Z_A} ce^{-\epsilon r}r^{n-1}\sin^{n-2}\theta_1\sin^{n-3}\theta_2\ldots\sin^2\theta_{n-3}\sin\theta_{n-2}drd\theta_1\ldots d\theta_{n-1}$$

$$= \int_0^R c_1 r^{n-1}e^{-\epsilon r}dr \int_0^\pi c_2\sin^{n-2}\theta_1 d\theta_1 \int_0^\pi c_3\sin^{n-3}\theta_2 d\theta_2\cdots\int_0^\pi c_{n-1}\sin\theta_{n-2}d\theta_{n-2}\int_0^{2\pi} c_n d\theta_{n-1}$$

where $\prod_{k=1}^{n} c_k = c$.

We note that this is a product of independent distributions, and thus the co-ordinates (radius and angles) can be selected independently. We also require that each integral sums to 1 to get valid probability distributions, so firstly we can calculate the constant c_1 using

$$\int_0^\infty c_1 r^{n-1} e^{-\epsilon r} dr = 1 \tag{6}$$

Using integration by parts we find

$$\int_0^\infty c_1 r^{n-1} e^{-\epsilon r} dr = \frac{c_1}{\epsilon^n}(n-1)!$$

And thus the integral in (6) becomes

$$\int_0^\infty \frac{\epsilon^n}{(n-1)!} r^{n-1} e^{-\epsilon r} dr = \int_0^\infty \frac{\epsilon^n}{\Gamma(n)} r^{n-1} e^{-\epsilon r} dr$$

which we recognise as the PDF of the Gamma distribution

$$f(x, k; \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

where $x \mapsto r$, $k \mapsto n$ and $\theta \mapsto \frac{1}{\epsilon}$.

Now, we also note that the remaining integrals (over the angles $\theta_1, \ldots \theta_{n-1}$) correspond to the Jacobian for the unit n-sphere. In other words, this is equivalent to the problem of selecting a point uniformly over the surface of the n-sphere.

Bibliography

[1] Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914 ACM. (2013).

[2] Chatzikokolakis, K., Andrés, M. E., Bordenabe, N. E., and Palamidessi, C. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102 Springer. (2013).

[3] Kroese, D. P., Taimre, T., and Botev, Z. I. *Handbook of monte carlo methods*, volume 706. John Wiley & Sons, (2013).

[4] Marsaglia, G. et al. Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, 43(2):645–646 (1972).

[5] Mustard, D. Numerical integration over the n-dimensional spherical shell. *Mathematics of Computation*, 18(88):578–589 (1964).

[6] Weggenmann, B. and Kerschbaum, F. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. *arXiv preprint arXiv:1805.00904* (2018).