A Novel Framework for Author Obfuscation using Generalised Differential Privacy

By

Natasha Fernandes

A thesis submitted to Macquarie University
for the degree of Master of Research
Department of Computing
9 October, 2017

MACQUARIE
University
SYDNEY·AUSTRALIA

Typeset in LaTeX 2_ε.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Natasha Fernandes

ACKNOWLEDGEMENTS

I am deeply grateful to my supervisor, Prof. Annabelle McIver, and my assistant supervisor, A/Prof. Mark Dras, for their constant encouragement, words of wisdom and outstanding good humour throughout this project. This thesis reflects their dedication and support, for which I am humbly thankful.

I'd also like to express my thanks to Tim Chard for his programming prowess and assistance with the experimental work in this thesis.

Finally, to my friends and family who supported me throughout the year. In particular, to my husband Anthony, who helps me to believe in myself, and to my beautiful boys, Ben, Jack and Tom, who remind me to approach every day with joy and wonder. Thank you for your love and your patience especially during the writing of this thesis.

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. Another approach to privacy, known as *differential privacy*, is advocated in the literature for its strong privacy guarantees. However, differential privacy has been dismissed as an option for text document privacy due to its design around the release of aggregate statistics, and its dependence on notions of 'adjacency', neither of which apply to text document privacy. In addition, differential privacy does not permit the release of individual data points as required for text document publishing. However, a new approach to privacy known as *generalised differential privacy* extends differential privacy to arbitrary datasets with no notion of adjacency, and permits the private release of individual data points. In this thesis, we show to apply generalised differential privacy to author obfuscation, drawing inspiration from the example of geo-location privacy, and utilising existing tools and methods from the stylometry and natural language processing literature.

Contents

List of Figures

List of Tables

1 INTRODUCTION

Satoshi Nakamoto, the pseudonymous creator of bitcoin, eluded identification for many years, using encryption and obfuscation techniques to mask his whereabouts and identity. Recently, his identity was allegedly established by the NSA using techniques from *stylometry*, the study of stylistic characteristics of writing [49]. The NSA used known texts by the bitcoin creator, and analysed the frequency of word use for common words across different sections of these texts. These word frequencies were used as a fingerprint to compare the bitcoin creator's writing against trillions of emails collected from different individuals until a match was found. Nakamoto's anonymity was by choice, however his request for privacy was not able to withstand the efforts of an organisation with access to large quantities of data. Importantly, whilst the problem of text de-anonymisation has received significant attention in the stylometry community, the reverse problem of text anonymisation has not, and there have been limited attempts at privacy in this domain. This raises the question: could a robust privacy approach to text anonymisation have protected Nakamoto's anonymity?

The problem of obscuring, or obfuscating, the authorship of a piece of writing is important for individuals who wish to protect their identity, such as whistleblowers or those wishing to release information without fear of reprisal. Freedom of speech and anonymity have long been recognised as rights by countries such as the US, and debates have extended into the right for anonymity in the online space. However, the right to anonymity is meaningless if writing style analysis can reveal an author's identity. The problem of text anonymisation, also known as author obfuscation, has only recently gained some attention in the research community. Whilst intuition says that authors could simply mask their own writing style by choice, research has found that authors can still be identified by their stylistic traits when attempting to write anonymously [44]. In addition, the tools used to identify authors of anonymised texts are becoming more accurate, as a result of significant research focused on the problems of author identification and author attribution. Meanwhile, little headway has been made into the problem of anonymising text documents. This means that authors who wish to write anonymously have few tools available to assist them in protecting their identity against adversaries armed with state-of-the-art author identification tools.

Author obfuscation is part of the text document privacy domain, which includes tasks such as redaction and sanitisation of sensitive documents. These tasks have applications in areas such as declassification for government agencies, or the release of healthcare documents in order to comply with federal privacy acts. Text document privacy in general encompasses tasks involving the removal or obfuscation of portions of text to protect sensitive information, whilst releasing the remainder of the

text untouched. Approaches to text document privacy in the literature revolve around the use of so-called *ad-hoc* privacy techniques. Privacy is usually evaluated *a posteriori*, on a particular dataset and against a particular adversary. These types of *ad-hoc* methods include approaches such as k-anonymity, which have been popular for some time due to their simplicity and their intuitive nature. However, a string of de-anonymisation attacks against public datasets, most famously against Netflix and AOL, has decreased confidence in the ability of *ad-hoc* methods to provide privacy protection against adversaries armed with auxiliary information [27]. This has caused a momentum shift in the research community away from *ad-hoc* privacy methods and towards privacy definitions which provide *a priori* guarantees.

A key development in recent years is a definition of privacy known as *differential privacy*. Differential privacy, introduced by Dwork, McSherry, Nissim & Smith [23], is rapidly becoming a consensus definition for privacy in the literature. Its strong mathematical foundations and provable privacy guarantees make it a compelling privacy proposition. Further, differential privacy does not rely on assumptions about the background knowledge of an attacker, and is therefore robust to linkage attacks which utilise auxiliary information. This sets differential privacy apart from previous definitions of privacy, and creates a new benchmark against which competing privacy definitions are examined. Differential privacy reframes privacy in terms of the participation of an individual, permitting information to be learned by an adversary provided the knowledge gained is independent of the presence or absence of the individual in the dataset. Whilst this definition has a natural interpretation for statistical databases, it is not obviously transferable to domains involving semi-structured or unstructured datasets, such as text document datasets. Moreover, differential privacy does not permit direct access to data, as required for private text document publishing. For these reasons, the consensus in the text privacy literature is that differential privacy cannot be applied to problems in the text document privacy domain.

However, a recent extension of differential privacy known as *generalised differential privacy* [12] has opened the way for applications of differential privacy to domains outside of structured databases. Generalised differential privacy permits the application of differential privacy to domains in which there is no natural notion of 'participation of an individual', instead drawing on the notion of distance between data points. Its definition can therefore be applied to arbitrary domains defined with a metric distance. Generalised differential privacy has been applied to the problem of geo-location privacy, which involves the release of geo-location co-ordinates in a differentially private manner [6]. This example provides a new way of looking at differentially private data release through the release of individual data points in an arbitrary domain. This raises the question of whether the same insight can be applied to author obfuscation. That is, can generalised differential privacy be applied to the release of a text document in such a way as to protect its authorship?

Contributions of this Thesis

In this thesis, we propose a novel approach to author obfuscation using the framework of generalised differential privacy. This represents the first application of differential privacy to a problem in the text document domain. We draw on existing notions of authorship from the stylometry and natural language processing literature which incorporate the use of distance measures between authors. Then we use the application of geo-location privacy as a blueprint for applying generalised differential privacy to the author obfuscation problem. Finally we consider the literature on mechanism design, and investigate the use of existing language technologies as noise-adding mechanisms to achieve author obfuscation.

The main contributions of this thesis are:

1. The development of a novel theoretical model for author obfuscation based on generalised differential privacy.
2. The experimental evaluation of existing technologies from the natural language domain to determine the feasibility of applying these to privacy mechanisms.
3. An identification of the gaps in the privacy literature which pose problems with respect to understanding or applying differential privacy to new domains.
4. An identification of deficiencies in natural language processing technologies which require further exploration in order to improve privacy mechanisms for text documents.

Organisation of this Thesis

This thesis is organised as follows: Chapter 2 provides background on the author obfuscation problem as well as prior work in this area. Chapter 3 explores privacy in general with a focus on differential privacy and text document privacy. Chapter 4 introduces generalised differential privacy and develops the theoretical framework for understanding author obfuscation. Chapter 5 contains experimental work used to investigate the feasibility of implementing author obfuscation using existing language tools. Chapter 6 concludes with a discussion of the key results obtained in this thesis as well as areas for future work.

2 AUTHOR OBFUSCATION

The purpose of this chapter is to clarify the author obfuscation task and highlight existing work on author obfuscation in the literature. We will first define the privacy goals for this thesis using the PAN 2016 author obfuscation task as a motivation. We then review prior work which highlights the gap in the literature for privacy in this field. Finally, techniques from stylometry and natural language processing will be introduced in order to provide some background on the approach taken later in this thesis.

2.1 Overview

The author obfuscation problem tackled in this thesis is motivated by the PAN 2016 author obfuscation task. [1] PAN shared tasks are open to the research community and designed to encourage research in areas involving text document forensics, such as author identification and plagiarism detection. [2] These tasks are typically attempted by researchers in natural language processing (NLP) or in the stylometry domain. The PAN 2016 author obfuscation task was introduced due to a lack of research in the area, and its introduction has opened up the problem to the NLP and stylometry research communities. Within these domains, more emphasis has traditionally been put on the reverse task of author identification (and its various incarnations). The PAN 2016 author obfuscation task received only 3 entries which is representative of the lack of research in the field at present. This points to a clear gap in the literature for this research area.

2.1.1 PAN 2016 Task

As most of the research in this field stems from the introduction of the PAN 2016 task, it is helpful to understand the author obfuscation task and goals as defined by PAN. The author obfuscation task is defined as: *Given a document, paraphrase it so that its writing style does not match that of its original author, anymore*. [3]

This is a privacy task that requires modifying a document so as to protect the identity of its author. As with any privacy task, there is a utility requirement which is presented as part of the evaluation criteria. The PAN task is evaluated using 3 metrics:

- *Safeness* - that is, whether the original author can be identified from the obfuscated text. A suite of automated author verification tools is used to evaluate safeness.

[1] The PAN 2017 author obfuscation task is identical, but for the purposes of this thesis, the PAN 2016 will be referred to as it contains the original statement of the problem.

[2] http://pan.webis.de/tasks.html

[3] http://pan.webis.de/clef17/pan17-web/author-obfuscation.html

- *Soundness* - that is, whether the obfuscated text semantically entails the original text. This is assessed by manual peer review.
- *Sensibleness* - that is, whether the obfuscated text is grammatically correct. This is also assessed by manual peer review.

This task motivates the definition for author obfuscation presented later in this chapter.

2.1.2 Prior Work

There has been relatively little written about the author obfuscation task, although there has been significantly more interest in the reverse task of author identification. Much work in this area has come out of stylometry, a field dedicated to the statistical analysis of the features of writing style which can be used to characterise authors. Early work in stylometry was done by Mosteller & Wallace [48] who famously determined the authorship of the Federalist Papers using stylometric techniques available to them in the 1960's. Mosteller & Wallace noticed that particular authors could be distinguished by their use of so-called *function words*, which are non-content words such as prepositions, pronouns and particles. Modern stylometry has expanded the set of distinguishing features from function words to a range of lexical, syntactic, semantic and document level features [60]. State-of-the-art stylometric methods can identify authors with accuracy better than 90%, and have been used on authorship sets containing tens of thousands of authors [44].

One of the earliest works on author obfuscation is the work by Kacmarcik & Gamon [31]. Their approach looked at the stylometric features used by author identification methods, in particular the frequency patterns of particular words. They changed the frequencies of these words to an 'average' level, as determined by a known set of K documents selected to obfuscate against. This approach provided safety against an attacker armed with a specific author identification method and attempting to identify a document from amongst the K possible documents. However, the approach did not attempt to output a new document, instead calculating the probability of successful authorship identification using hypothetically calculated term frequencies.

This approach has been extended to the creation of obfuscated documents via a tool known as Anonymouth [44]. Anonymouth is a semi-automated author obfuscation tool built on top of the author identification tool JStylo, also by the same authors. [4] JStylo uses a technique called the Writeprints method to identify authorship features at high granularity with respect to a particular document set [2]. The Anonymouth tool uses these author-identifying features to suggest edits to the writer until Anonymouth is satisfied that the document has been successfully anonymised. This tool represents the state-of-the-art for author obfuscation at present.

[4] JStylo can be downloaded from https://psal.cs.drexel.edu/index.php/JStylo-Anonymouth

The author obfuscation task was introduced into the PAN 2016 competition in order to encourage research in this area. As mentioned above, the task only received 3 entries, compared with 10 entries received for author identification and 22 entries for author profiling. The 3 entries in author obfuscation used very different approaches, which is to be expected in a new domain with little pre-existing literature. One approach employed machine translation to perform obfuscation [34]. This method involved using an automated translation tool to translate a document from English to German to French and back to English. This technique was not successful at maintaining the semantics of the original document, nor at rendering text that was readable. A second approach, based on stylometry, utilised word frequencies to identify authorship for both the training document set and the document to obfuscate [43]. Word frequencies are popularly used in stylometry as they have been shown to characterise authors, in part due to their consistency of use across genres. Words in the document were systematically changed using synonym substitution until the word frequencies matched the training data, thus obscuring the (author-identifying) word frequencies in the original document. However, these substitutions were deliberately limited to at most one word per sentence in order to preserve the sensibleness of the resulting text. This limited the impact of the obfuscation technique. The final approach looked at a range of stylometric features such as average sentence length and punctuation to word ratio, calculating 'average' metrics for these features over the training set [46]. The document was then randomly modified using strategies such as synonym substitution, sentence splitting and paraphrasing to change its metrics to more closely match the metrics for the training set. This was performed for all documents in the training set; in other words, the method reduced all document metrics to the (approximately) same average value. This approach aimed to reduce the accuracy with which existing methods could identify authorship, and was the most successful entry at PAN 2016 [53].

It is clear from studying the literature that the approaches used for author obfuscation to date use *ad-hoc* methods for privacy, and aim at preserving the sensibleness, or grammatical correctness, of the text. As this is a developing field, there has not yet been an emphasis on the privacy properties for these obfuscation techniques, nor an interest in examining the privacy guarantees that these techniques provide. This motivates the goal of this thesis.

2.1.3 Privacy Goal for this Thesis

The focus for this thesis is on author obfuscation as a *privacy* task rather than a language task, which represents a key shift in emphasis compared with the work in the literature to date. The PAN 2016 task will be used as the motivating task for this thesis, however some modifications will be made in order to reduce the complexity of the language component of this problem. As a first simplifying step, we will only require that the output document preserves the *topicality* of the original document. This sort of simplification is reasonably standard in the NLP literature and has been used to demonstrate

the feasibility of new approaches in the domain [16]. This means that the sensibleness criteria can be removed, as well as the semantic entailment requirement for the obfuscated text. This simplification does not destroy the privacy aspect of the task, as the connection between topicality and authorship identification has been established in the stylometry literature, and is assumed to be a first step towards author privacy whilst preserving the full document semantics. [5]

As a second step, the privacy goal will be reframed as an anonymisation goal, to take into account factors (other than writing style) which can contribute to the identification of an author. In addition, the utility aspect of the task (the preservation of topicality) will be explicitly included. This motivates the consideration of privacy and utility together. Thus the author obfuscation task for this thesis will be framed as:

Definition 1. *(Author Obfuscation) Given a document, obfuscate it in such a way as to prevent identification of its author whilst preserving the topicality of the original text.*

Obfuscated texts will be evaluated using the following metrics:

- *Safeness* - that is, whether the original author can be identified from the obfuscated text. This will be evaluated using a single state-of-the-art author verification tool rather than a suite of automated verifiers.
- *Soundness* - that is, whether the obfuscated text preserves the topicality of the original text. This will be evaluated using a machine learning classifier trained on a set of documents labelled by topic.

From a privacy perspective, it is helpful to also consider the role of the adversary. The adversary in this task aligns with the *safeness* evaluation metric; that is, an author verifier trained to identify authors using a particular training set of documents.

2.2 Stylometry and NLP

In this section we introduce stylometric and NLP methods used for author identification. We focus on a particular method based on *character n-grams* which will be used later in this thesis to develop our understanding of author obfuscation.

[5] Although some authors consider writing style and topicality to be orthogonal, topic-specific words have been shown to be identifying for documents within a genre.

2.2.1 Overview

Stylometry is the study of stylistic traits in documents which are characteristic of particular authors. Stylometric analysis involves the manual selection of 'features' of a text which are then used by classifiers to determine the most likely author of a document. Stylometric features can act as a fingerprint for an author in many domains. Tweets, emails, and even programming code can be imprinted with enough stylistic traits to identify their authors. Stylometric analysis was famously used to identify the authorship of the disputed Federalist Papers, a set of anonymous papers written between 1787 and 1788 designed to convince readers in New York to ratify the constitution [48]. It has more recently been used to identify J. K. Rowling as the most likely author of *The Cuckoo's Calling* [30], and was reportedly used to identify up to 80% of anonymous forum users through their posts in an underground forum [51]. Its application to tweets has been used to identify a number of possible speech writers used by the Prime Minister of Pakistan [32], as well as President Trump's use of a second 'tweeter' for presidential announcements [54].

Although the task of author identification traditionally falls under the study of stylometry, the NLP community has more recently become interested in this problem. NLP techniques for authorship analysis typically make use of machine learning methods to classify documents by author (using supervised methods) or to cluster documents with similar features (using unsupervised methods). Machine learning methods require vector representations of their inputs, thus documents are usually first transformed into *feature vectors*. These features can be the same ones used in stylometric analysis, or they can be learned by the particular machine learning method. Although machine learning methods can be high-performing, it is usually difficult to interpret their learned features. For example, a character-based neural network was used in the PAN 2015 task on authorship attribution [6] and represented the best performing method for that task [8]. However, neural networks learn complex features which are notoriously difficult to interpret. Thus, for this thesis we will choose a stylometric feature set for author attribution which is considered state-of-the-art.

There are three types of stylometric features employed in the literature: 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"

[6] The terms authorship identification, authorship verification and authorship attribution have slightly different meanings but will be used interchangeably in this thesis.

would be ‘The’, ‘her’, ‘ere’, ‘re_’, ‘e_it’, ‘it_’, ‘t_i’, ‘_is’ [7]. The choice for the value n is determined by the document set and the language in use; larger values for n (4 or greater) are typically better at capturing contextual information whilst smaller values can capture more typical stylistic features such as prepositions and other non-content words [60]. Character n-grams are considered the current state-of-the-art in authorship identification and this is the representation chosen for this thesis.

2.2.2 A Character n-gram-Based Approach

As mentioned earlier, character n-grams have been shown to be effective features for authorship attribution and represent the state-of-the-art in this area. The intuition behind using an n-gram rather than a whole word is that n-grams capture extra features such as misspellings, punctuation and common word stems, all of which can be characteristics for an author. For example, the words ‘task’ and ‘tasked’ share the same first 4-gram ‘task’, however a word-based approach classifies these words as separate, unrelated features. Character n-grams also capture both stylistic and content based features of documents [36], which make them useful for our obfuscation task.

A particular character n-gram-based method of interest is the one developed by Koppel et al. [36], which we refer to as the *Koppel Method* in this thesis. The authors utilise character 4-grams to classify authorship on a document set consisting of blog posts from thousands of authors, and achieve in excess of 90% precision with 42% coverage for a 1000-author dataset. This is considered one of the best-performing approaches for large author sets, as many existing authorship verification tools are evaluated on only a handful of authors [36]. Current state-of-the-art author verification tools use the Koppel Method as a core for their algorithms [35], including the winning entries in the PAN 2013 and 2014 authorship verification tasks [59, 61]. We will use this method as the author verifier for the *safeness* evaluation in the author obfuscation task.

2.3 Summary

The author obfuscation task is a relatively new problem in the text document privacy space. The few approaches to this problem use *ad-hoc* methods to achieve privacy, and it is unclear what privacy guarantees these methods offer. There have been no attempts to analyse privacy for these approaches, aside from evaluating their performance against author identification methods. In this thesis we will provide a privacy-oriented approach to author obfuscation. The stylometric approach of Koppel et al. [36] will be used to evaluate the *safeness* aspect of the author obfuscation task proposed for this thesis. We will also see that character n-gram features provide a useful vector representation for documents which will be used when defining author obfuscation in terms of generalised differential privacy.

[7] Note that we use ‘_’ to represent spaces.

3 AN EXPLORATION OF PRIVACY

The purpose of this chapter is to provide an overview of privacy with a focus on differential privacy and the text document space. We begin with a background to privacy, culminating in the introduction of differential privacy which represents the dominant definition of privacy in the literature. We see that differential privacy offers a number of key features which make it a compelling choice for privacy. We then explore privacy in the text document domain, including a further exploration of applications to machine learning and information flow approaches. We find that text document privacy is still a developing field and there is no consensus definition for privacy in this domain. In particular, there are no applications of differential privacy in this area.

3.1 A Brief History of Privacy

Data privacy is a well-established research area, with its origins in statistical data disclosure. Early work on privacy stemmed from concerns over the release of statistics from census datasets, and whether these could be traced back to individuals. Dalenius is usually credited with the first definition of privacy, which is often paraphrased as *access to a statistical database should not enable one to learn anything about an individual that could not be learned without access* [22]. This definition, however, is not achievable, as hypothesised by Dalenius himself [17] and demonstrated by Dwork et al [23]. They reasoned that information about an individual can be inferred by access to a statistical database using auxiliary information about the individual *even if the individual is not in the database*. For example, an adversary who knows only that Terry Gross is 2 inches shorter than the average Lithuanian woman, can determine Terry's exact height using a statistical database that returns the height of the average Lithuanian woman. This research concluded that not only is Dalenius' definition unachievable, but it does not accurately capture what is meant by privacy. This led to a new definition of privacy known as *differential privacy*.

Until these insights, privacy definitions continued to use intuitive notions of privacy which relied on assumptions about the auxiliary information available to an attacker. Research turned to anonymisation of datasets, which requires removing or modifying *all* information which can be used to identify an individual. Anonymisation is frequently confused with the weaker notion of de-identification, which simply involves removing explicit personal identifiers such as names, addresses and dates of birth. De-identification is popularly seen as a safe method of anonymisation; however, it has long been known that de-identification does not guarantee anonymity [62]. De-identification methods are vulnerable to linkage attacks because of the existence of so-called 'quasi-identifiers'[1] which can be

[1]A quasi-identifier is a tuple of attributes which taken together are uniquely identifying for individuals in a dataset.

used together to uniquely identify individuals [18]. For example, a study by Sweeney [63] found that 87% of individuals in the United States could be uniquely identified from their date of birth, zip code and gender. In 2002, Sweeney identified the medical data of the Massachusetts governor in de-identified hospital records by linking his quasi-identifiers with information freely available in a voter database [64]. This research led to a framework for anonymisation known as *k-anonymity*.

K-anonymity is a popular and intuitive anonymisation model that protects privacy in a dataset by ensuring that for every row containing a particular quasi-identifier there are at least $k-1$ other rows containing that same quasi-identifier. However, k-anonymity has some shortcomings which can cause the confidential attributes for an individual to be revealed. For example, a k-anonymous medical dataset may reveal that all individuals with a particular quasi-identifier have cancer, thereby disclosing the sensitive information of every individual in the group [42]. Alternative anonymity models have been proposed which aim to correct the flaws in k-anonymity; these include *l-diversity* [42] and *t-closeness* [40]. However, all of these models share the same flaw as Dalenius' first privacy definition: they rely on assumptions about an adversary's background knowledge. These assumptions became the focus of attention after a number of high-profile privacy breaches.

In 2006, Netflix released anonymised movie ratings data for 500,000 of its users as part of a competition designed to encourage the development of machine learning algorithms for the Netflix recommendation system. However, a team of researchers de-anonymised the Netflix data by correlating it with data from another online movie database, resulting in a lawsuit against Netflix [50]. In the same year, AOL released anonymised search engine logs for use by researchers, only to have an individual re-identified by writers at the New York Times based on her search queries [9]. More recently, it was shown that only 4 data points are required to identify 90% of individuals from a dataset of anonymised credit card transactions [19]. These examples highlight the difficulties of providing a privacy guarantee for data which has been anonymised using an 'intuitive' notion of privacy. There is a trend in the literature to move away from such *ad-hoc* models of privacy and towards theoretical models which can provide *a priori* privacy guarantees. Differential privacy is one such model.

3.2 Differential Privacy

As mentioned earlier, research by Dwork et al [23] provided an important insight about privacy, namely that information about an individual can be learned even when that individual is not present in the dataset. This led them to move away from definitions of privacy which focus on what can be learned from a dataset, and towards a definition which aims to protect individuals within a dataset. Differential privacy, then, is concerned with the harm caused to an individual by *participation* in a dataset. Differential privacy promises that the presence or absence of an individual in a dataset will not

significantly change the result of a statistical query over that dataset. In other words, similar datasets should produce results which are indistinguishable. This prevents an adversary from being able to infer whether or not an individual's data was in the dataset. Note that it does not prevent sensitive information from being learned about the individual, as this could happen regardless of whether the individual was in the dataset or not, as the Terry Gross example demonstrates. Nor does it limit the amount that is learnable from the dataset, as this is precisely what makes datasets useful. The privacy guarantee ensures that nothing more is learned about an individual by their presence in the dataset, than would have been learned in their absence.

Differential privacy has a number of key features which have contributed to its popularity. Firstly, its mathematical formalisation permits proofs of its privacy guarantees, and has been used to prove properties such as composability and post-processing invariance [24]. Secondly, its definition is independent of auxiliary information, making differential privacy robust to the sorts of linkage attacks for which k-anonymity is vulnerable. Finally, differential privacy defines privacy in terms of risk of disclosure, which importantly recognises that privacy is not simply 'true' or 'false', and this allows the amount of privacy for a dataset to be parametrised. These features make differential privacy a compelling choice for privacy practitioners who wish to provide *a priori* privacy guarantees.

The definition of differential privacy depends on the notion of the *participation* of an individual in a dataset. This notion is formalised as follows: given a universe of values $\mathcal{V}$ and a dataset $\mathcal{V}^n$ consisting of n individuals, differential privacy promises that the output of a randomised algorithm K on a dataset $x \in \mathcal{V}^n$ will be (almost) indistinguishable from the output of K on an *adjacent* dataset $x' \in \mathcal{V}^n$. Adjacency is defined as 'differing in the value of a single individual', which has a natural interpretation for structured datasets of individuals; here, if x and x' are adjacent datasets, then $d_h(x, x') = 1$ where d_h is the Hamming distance defined on $\mathcal{V}^n$. The notion of protecting participation of an individual provides privacy for the individual's values in the dataset; an adversary cannot determine, after observing an output, whether it originated from a dataset x containing the individual's values, or the adjacent dataset x' not containing the individual. Thus the individual's presence or absence make no difference to the conclusions drawn by such an adversary.

Differential privacy can be formalised in the following way [24]:

Definition 2. *(Differential Privacy) A randomized algorithm $\mathcal{K}$ satisfies ϵ-differential privacy if for all $\mathcal{S} \subseteq$ Range($\mathcal{K}$) and for all $x, x' \in \mathcal{V}^n$ such that $d_h(x, x') = 1$:*

$$Pr[\mathcal{K}(x) \in \mathcal{S}] \leq e^{\epsilon} Pr[\mathcal{K}(x') \in \mathcal{S}],$$

where the probability space is over coin flips of the mechanism $\mathcal{K}$.

Differential privacy is parametrised by a variable ϵ, commonly referred to as the *privacy budget*. Intuitively, smaller values for ϵ correspond to stronger privacy.

The requirement for the datasets x, x' to be adjacent is important, as a mechanism whose outputs are required to be indistinguishable for *any* pairs of datasets x, x' is useless; nothing at all can be learned from the output of such a mechanism. In addition, the interpretation of *adjacency* is important for applying differential privacy to a privacy task. This corresponds to the notion of an 'individual' in the datasets, and represents the *granularity* of the privacy guarantee. For example, a social network dataset may be modelled as a graph structure with nodes representing individuals and edges representing relationships. Applying differential privacy at the granularity of nodes protects the presence or absence of the individual within the social network. Alternatively, differential privacy at the granularity of edges protects the relationships within the social network. The level of privacy to apply depends on the privacy and utility requirements for the system. Weaker (edge) privacy may be preferred to ensure better utility of queries over the data.

3.2.1 Privacy-Utility Trade-Off

It has been well documented in the literature that the implementation of differential privacy comes with a trade-off in utility [4, 22, 28]. Utility refers to the 'usefulness' of the data; that is, how much can still be accurately learned from the data after the application of privacy. For example, applying no privacy at all results in perfect utility, as no information is lost; on the other hand, returning a random result from the dataset (independent of the query), results in perfect privacy but no utility, as no useful information can be learned from this output. Some criticisms of differential privacy stem from the resulting loss of utility from the dataset, which is often of secondary concern and considered less important than protecting privacy [22]. In one extreme case, it was reported that the use of differential privacy to protect patient privacy in a medical experiment would have resulted in a number of deaths from over-medicating, due to the noise introduced by the differentially private mechanism [26]. Much recent work has focussed on measuring this privacy-utility trade-off, and in the development of mechanisms which produce results with better utility for the same measure of privacy [1, 52]. We consider the measurement of utility important for practical applications of differential privacy, and will use both utility and privacy metrics for evaluating results in this thesis.

3.2.2 Limitations of Differential Privacy

Differential privacy has a number of known limitations which are of particular relevance for its application to text document privacy.

Firstly, the definition of differential privacy does not naturally lend itself to use in domains involving semi-structured or unstructured datasets, in which there is no natural notion of 'adjacency', nor an

interpretation for 'individuals' in the dataset. It is unclear how to translate differential privacy into a meaningful privacy definition for these domains.

Secondly, differential privacy's guarantee degrades with every access made to a dataset, as each access leaks more information about the dataset and therefore erodes the privacy budget. Differential privacy is consequently more commonly used in 'interactive' mode, in which direct access to the dataset is restricted via the use of queries whose (noisy) results are released. The number and types of queries is necessarily restricted in order to avoid consumption of the privacy budget. This has generally restricted the use of differential privacy to data mining scenarios rather than data publishing scenarios, as required for most text document privacy applications. Some authors claim that the use of privacy in these scenarios causes a degradation in utility rendering its use impractical [56].

Finally, the use of differential privacy for data publishing scenarios is typically enabled via the creation of synthetic datasets [7, 10, 29]. However, differential privacy does not permit publishing an individual data point; in particular, it would require that any change in the value for the data point should not affect the output, thus destroying any utility [6].

3.2.3 Mechanisms for Differential Privacy

Much work on differential privacy in the literature focuses on the implementation of mechanisms which satisfy the privacy definition. Of particular interest in this thesis are mechanisms based on the Laplace mechanism [24]. This is a popular mechanism for differential privacy, and is implemented through the addition of noise drawn from a Laplace distribution. The amount of noise to add is determined by both ϵ and the *sensitivity* of the query over the data. The sensitivity measures how much the result from the query can vary with the presence or absence of an individual. For example, a counting query has sensitivity of 1, since the presence or absence of a single person can change the result of a count by at most 1. However, a maximum value query (such as 'return the maximum age') has a much higher sensitivity, as the presence of an individual could radically change the result of such a query. Intuitively, the more noise that is applied, the less useful is the output from the dataset. We will see an adaptation of Laplacian mechanisms (Section 4.1.3) which replaces the notion of *sensitivity* with the notion of *metric distance*.

Also of note is the exponential mechanism, introduced by McSherry and Talwar [45], which is useful for applying differential privacy to domains in which the output is discrete rather than continuous, or where the application of noise can completely destroy the utility of the data. This mechanism uses a score function defined over the inputs and outputs of the system, and probabilistically chooses higher scoring outputs with exponentially higher likelihood than less useful scores.

3.3 Text Document Privacy

We now consider approaches to privacy in the text document domain. Text document privacy in the literature usually refers to the tasks of sanitisation and redaction of text documents, which are two commonly employed approaches used to hide sensitive textual information. Redaction refers to the removal or 'blacking out' of portions of text which contain sensitive information; this technique is commonly used for declassifying sensitive government documents. Sanitisation refers to the replacement of sensitive terms with more general terms which obscure the sensitive information. For example, the term 'HIV' may be replaced with the more general 'disease' in a healthcare dataset to protect knowledge of an individual's HIV status.

Few models have been proposed to implement automated redaction or sanitisation of text documents. Cumby & Ghani [16] develop a document sanitisation system inspired by the model of k-anonymity for structured datasets, which they term *k-confusability*. Under k-confusability, sensitive terms are generalised so that a classifier attempting to guess the sensitive term finds at least $k - 1$ other terms which are equally likely. Anandan et al. [5] develop a text privacy definition (also inspired by k-anonymity) known as *t-plausibility*, which requires that the output document from a t-plausible mechanism using an ontology could have been produced from at least t possible input documents. Sanchez & Batet [57] propose a text privacy definition known as *C-sanitisation* which requires that a C-sanitised document cannot disclose any more semantics of the sensitive text than is otherwise disclosed by using (non-sensitive) generalisations of the text. These approaches rely on the identification of sensitive portions of text and do not require that the remainder of the document be sanitised. Like k-anonymity, they could be vulnerable to linkage attacks by adversaries with knowledge about the untreated text. In addition, they do not quantify the privacy guarantee for their mechanisms, which makes it more difficult to reason about their privacy properties.

In summary, approaches to privacy in redaction and sanitisation are currently ad-hoc, as this is still a developing research area. Much focus is on the use of language tools and maintaining sufficient semantics in the resulting text documents. As yet there have been no approaches specifically focused on privacy models providing strong privacy guarantees, and in particular there have been no attempts to apply differential privacy in this domain.

3.4 Other Approaches to Privacy

In this section we explore other approaches to privacy which were considered in the context of text document privacy, namely machine learning approaches and privacy approaches using information flow.

3.4.1 Machine Learning

Machine learning is a popular technique for creating predictive models based on very large datasets. Most machine learning approaches use supervised methods, which involves the creation of a 'training' dataset used by the machine learner to build a model of the data. The machine learned model is then applied to a 'test' dataset to evaluate its predictive accuracy on unseen data. The success of these learners typically hinges on the size of the training set used; neural networks, for example, require datasets in the order of millions of data points in order to build a sufficiently accurate model. However, the popularity of machine learning lies in the diversity of application of these models to real world problems. Machine learning tasks have been applied to various problems including face recognition and document translation. Recent work by Google has seen machine learners used to detect cancer tumours from pathology images with better accuracy than a trained pathologist [41]. These sorts of applications are driving the growth of machine learning techniques in industry.

The use of machine learning on datasets which contain highly sensitive data raises questions about privacy and in particular whether machine learning mechanisms leak information on datasets that they have been trained on. The complexity of many machine learning algorithms makes it difficult to analyse the internal workings of machine learners. However, adversarial attacks against machine learners have been used to reconstruct data from training sets, demonstrating that even complex learners such as neural networks leak sufficient information to perform a reconstruction attack [25]. This has created a need for the development of privacy mechanisms for machine learning algorithms.

Differential privacy has a natural application to machine learning, as both tasks emphasise the release of learned statistics over the data without depending too much on the values of individuals. This has resulted in a large amount of work dedicated to building differentially private machine learning algorithms. (See [1, 11, 13, 14] for examples). In the machine learning literature, emphasis has been placed on the application of noise to the internals of the algorithm in order to satisfy differential privacy, as well as measurements of utility through experiments designed to measure the accuracy of the algorithm on some dataset.

However, differential privacy in the machine learning literature is interested in the protection of the training dataset against adversarial attacks [1, 52]. Our interest for this thesis is on privacy guarantees for released data, such as obfuscated documents. As such, the applications of differential privacy in machine learning were not relevant for our task.

3.4.2 Information Leakage

Another avenue for exploration was the use of information leakage measures to provide privacy metrics in terms of information gained by an adversary. Whilst differential privacy cautions against such

approaches, information leakage is well-established in the literature and can be used in conjunction with differential privacy to determine how much information is leaked when a differentially private release occurs [3]. Information leakage measures are not domain specific, so these models can be applied to problems in the text document privacy domain.

Of particular interest were information theoretic approaches with connections to differential privacy and applications to text document privacy. We only found one method of interest, which was the work of Calmon & Fawaz [20] who model the privacy problem using rate-distortion theory. The authors provide an information-theoretic framework which measures the information learned by an adversary when a user privately releases information. Their framework can be applied to unstructured data domains, and they provide an example of the release of social networking data to a recommender system. We thus considered this an approach of interest for text document privacy tasks. The authors provided a link between differential privacy and their definition of information privacy. In particular, they claim that information leakage is *unbounded* under differential privacy. This appears to contradict the result by Alvim et al. [3], which proves that ϵ-differential privacy implies a bound on Shannon entropy leakage and also min-entropy leakage.

However, we show that Calmon & Fawaz's proof has some subtle flaws which cast doubt on their result. They claim that ϵ-differential privacy does not provide any guarantee on the information leakage. This claim is proven via the construction of a particular dataset in which information leakage is lower bounded for a given ϵ-differential privacy guarantee. We refer to Thereom 4 in [20], which says that for every ϵ and δ there exists a privacy mapping which is ϵ-differentially private but leaks at least δ bits on average. We make the following observations:

1. They construct a dataset which is constrained in such a way that a counting query can only return values in multiples of k. This means that the sensitivity of the counting function is k, and thus the parameters of the Laplace noise used in the proof need to be adjusted to use the sensitivity k. We note that this only changes the result slightly.

2. The authors show that leakage can be made arbitrarily large when the size of the dataset, n, is increased, for fixed ϵ. However, this is exactly what we would expect; the differential privacy promise is that confidential information of every individual in the database is not compromised, and hence the number of bits in the 'secret' is of the order of the size of the database, that is, n. Thus, this does not demonstrate that the leakage is unbounded in general; this would require fixing n and varying ϵ, as done in [3].

3. The authors show that the information leakage from the system is bounded below, as given by the following inequality:

$$I(S;U) \geq (1 - e^{-\frac{k\epsilon}{2}}) \log(1 + \frac{n}{k}) - 1$$

It is worthwhile checking whether this bound is in fact significant. Substituting $k = 1$, $n = 10000$ and $\epsilon = 1.0$ yields 4.2 bits of information leaked from a database of 10000 individuals (hence secret at least 10000 bits). Substituting an even larger value of $n = 100000$ yields 5.5 bits of information leaked. This level of information leakage is tolerable, and only increases logarithmically as n increases. Thus we do not consider that this lower bound demonstrates arbitrarily large leakage, as claimed.

The use of information flow techniques was not explored further, however this approach remains of interest for future investigation.

3.5 Summary

Our exploration of privacy found that there are a variety of approaches used in the text document space, however there are no clear consensus definitions for privacy in this domain. Methods in text document privacy are currently *ad-hoc* and typically incorporate notions from k-anonymity. Applications of privacy in machine learning and information flow were investigated, however no clear approaches were found that could be used in the text document privacy space.

Differential privacy is clearly the preferred choice for privacy over *ad-hoc* methods. Differential privacy has provable *a priori* privacy guarantees and is robust to future attacks by adversaries armed with auxiliary information, which has contributed to its popularity in the literature. However, its use has been limited to structured datasets due to limitations around differentially private data publishing and the difficulty of re-interpreting its privacy promise for unstructured datasets. Importantly, differential privacy does not permit the private release of datasets consisting of individual data points. These limitations have been used to argue that differential privacy cannot be applied to problems in the text document domain [16, 56]. As such, there are currently no applications of differential privacy in the text document domain.

However, in the next chapter we introduce the notion of generalised differential privacy, which permits the application of differential privacy over arbitrary domains, and gives the intuition needed to enable a definition of privacy for author obfuscation with differential privacy's guarantees.

4 THEORETICAL FRAMEWORK

The purpose of this chapter is to set out a novel theoretical framework for author obfuscation based on generalised differential privacy. We first introduce generalised differential privacy and then the example of geo-location privacy, known as *geo-indistinguishability*, which involves the private release of individual data points. Next, we show how the author obfuscation task can be framed in the same way as the geo-location privacy problem. However, we outline a number of differences which make the application to author obfuscation non-trivial. Finally we create a new privacy definition based on a word-based document representation which we propose can be used to provide author privacy.

4.1 Generalised Differential Privacy

Generalised differential privacy, introduced by Chatzikokolakis et al. [12], is a recent extension of differential privacy designed for use over arbitrary domains in which there is no natural notion of 'adjacency'. Recall from Definition 2 that standard differential privacy is defined in terms of adjacent datasets having Hamming distance 1. Generalised differential privacy extends the definition of differential privacy to arbitrary *metrics*. The metric distance $d_{\mathcal{X}}(x, x')$ between two datasets x, x', also known as the *distinguishability* between the datasets, is used to parametrise the amount of privacy. This metric replaces the Hamming metric for adjacency in differential privacy, and thus provides a way of identifying 'close' (if not adjacent) elements of a dataset.

Generalised differential privacy can be understood using the notion of indistinguishability between secrets. If the secrets x and x' are close together with respect to the metric $d_{\mathcal{X}}$, we say that they are *indistinguishable*. Thus any output from a privacy mechanism should occur with similar likelihood regardless of whether the input was x or x'. Conversely, if x and x' are distant with respect to $d_{\mathcal{X}}$, and therefore highly distinguishable, the output distributions should be very different. Thus it would be easier to guess whether an output came from x or x', which corresponds to a lower level of privacy for x and x'. This is the same intuition that we get from standard differential privacy, which uses the notion of 'adjacency' in place of indistinguishability to determine the privacy guarantee.

Generalised differential privacy is formalised as follows [12]:

Preliminaries

In the definition to follow, $\mathcal{X}$ and $\mathcal{Z}$ are sets, $\mathcal{F}_{\mathcal{Z}}$ is a σ-algebra over $\mathcal{Z}$ and $\mathcal{P}(\mathcal{Z})$ is the set of probability measures over $\mathcal{Z}$. $K : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ is a probabilistic function known as a *mechanism*. $d_{\mathcal{P}}$

is a metric on $\mathcal{P}(\mathcal{Z})$ defined as:

$$d_{\mathcal{P}}(\mu_1, \mu_2) = sup_{Z \in \mathcal{F}_{\mathcal{Z}}} \left| ln \frac{\mu_1(Z)}{\mu_2(Z)} \right| \qquad \forall \mu_1, \mu_2 \in \mathcal{P}(\mathcal{Z})$$

The convention is that if $\mu_1(Z)$ and $\mu_2(Z)$ are both zero then $\left| ln \frac{\mu_1(Z)}{\mu_2(Z)} \right| = 0$, and if only one of $\mu_1(Z), \mu_2(Z)$ is zero then $\left| ln \frac{\mu_1(Z)}{\mu_2(Z)} \right| = \infty$.

Definition 3. *(Generalised Differential Privacy) A mechanism $K : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ satisfies $d_{\mathcal{X}}$-privacy, iff $\forall x, x' \in \mathcal{X}$:*

$$d_{\mathcal{P}}(K(x), K(x')) \leq d_{\mathcal{X}}(x, x')$$

or, equivalently,

$$K(x)(Z) \leq e^{d_{\mathcal{X}}(x,x')} K(x')(Z) \quad \forall Z \in \mathcal{F}_{\mathcal{Z}}$$

Note that the ϵ parameter from differential privacy is omitted from this definition. Chatzikokolakis et al. [12] incorporate ϵ into the metric $d_{\mathcal{X}}$, noting that a scaled metric is also a metric. The authors also note that this definition can be reduced to standard differential privacy by substituting $\mathcal{X} = \mathcal{V}^n$ and $d_{\mathcal{X}} = \epsilon d_h$ (where d_h is the Hamming distance over $\mathcal{V}^n$).

4.1.1 Comparison with Differential Privacy

Recall that differential privacy protects the participation of an individual; if x and x' are adjacent datasets, then differential privacy guarantees that the output distributions from these datasets will be similar, regardless of whether the input was x or x'. Similarly, if x and x' are further apart (that is, differing in more than one individual), then differential privacy says that the privacy guarantee degrades according to the Hamming distance between x and x'. [1]

However, we note some key differences between differential privacy and generalised differential privacy which impacts our understanding of generalised differential privacy.

Firstly, the notion of datasets consisting of rows of individuals has been replaced by the notion of arbitrary domains containing no specific notion of 'individuals' within that dataset. This means that the intuition behind 'participation of an individual', which characterises differential privacy, is lost in generalised differential privacy. However, the authors capture some alternative intuition by considering privacy from the perspective of the adversary and, importantly, this intuition also applies to standard differential privacy. This intuition will be visited in Section 4.1.2.

[1] This is just an application of group privacy for differential privacy, which says that if the Hamming distance is k, then the mechanism satisfies $k\epsilon$-differential privacy. See Theorem 2.2 in [24].

Secondly, generalised differential privacy applies to the private release of secrets from the domain, in contrast with differential privacy which is used to release aggregate statistics over a dataset. Thus the database against which privacy is guaranteed in standard differential privacy corresponds with *individual secrets* in generalised differential privacy. Importantly, this means that the privacy parameter ϵ (which, although omitted from the above definition, will be used later to parametrise privacy) needs to be set for each released data point.

Finally, as already mentioned, the notion of adjacency between datasets has been replaced by a distance metric. This permits a stronger privacy guarantee for datasets which are closer together. This is because the values of the secrets determine the distance between the datasets (rather than the presence or absence of individuals) and hence the privacy guarantee is determined by the relative sizes of the secrets. This can be applied to structured datasets to provide stronger privacy than standard differential privacy by using an alternative metric to the Hamming distance to determine the distance between the datasets.

4.1.2 Characterisations of the Adversary

A key contribution of Chatzikokolakis et al. [12] is their drawing out of the characterisations of the adversary implied by differential privacy, and their equivalents for generalised differential privacy. These characterisations are useful in reasoning about the capabilities of an adversary, and also in providing some intuition about generalised differential privacy.

The first characterisation makes use of an arbitrary hiding function, $\phi : \mathcal{X} \to \mathcal{X}$, which replaces an point x with the point $\phi(x)$ before applying the mechanism K.

Characterisation 1. *(Generalised Participation of Individual) Regardless of side knowledge, the adversary's conclusions (captured by the posterior distribution) are similar (up to a factor ϕ) whether or not a hiding function ϕ was applied to the secret.*

As noted earlier (see Section 4.1.1), the notion of 'participation of an individual' is lost in generalised differential privacy. However this characterisation recaptures its meaning in terms of the distance between an individual x and its hidden version $\phi(x)$. This re-interprets the guarantee of differential privacy which says that any knowledge gained about an individual is independent of their presence or absence in the database. The choice of hiding function determines the granularity of the privacy guarantee, in a similar way that differential privacy's notion of 'adjacency' determines the granularity of its privacy guarantee.

Characterisation 2. *(Generalised Informed Adversary) An informed adversary who knows that the*

secret belongs to a neighbourhood N gains little information about the exact secret, regardless of prior knowledge within N.

This is an extension of differential privacy's promise that nothing more is learned about individual i by an informed adversary who knows every value except i's in the dataset.

These characterisations both importantly capture information about the adversary's gain of knowledge after observing an output from the mechanism K. They also provide some key insights regarding the difference between standard differential privacy and generalised differential privacy. However, we will see another notion introduced in geo-indistinguishability which we believe encapsulates these insights, namely the notion of 'l-privacy within radius r'. This is the notion that we will use throughout this thesis to understand how differential privacy applies to author obfuscation.

4.1.3 Mechanisms for Generalised Differential Privacy

In order to take advantage of the extensive literature on mechanisms for differential privacy, it would be useful if these could also be applied in the generalised differential privacy setting. The authors show that the Laplace mechanism, considered as an extension of the Exponential mechanism, always satisfies $d_{\mathcal{X}}$-privacy. We notice that this occurs specifically when $d_{\mathcal{X}}$ is a metric, thus justifying the metric requirement for generalised differential privacy.

The Laplace mechanism for generalised differential privacy is defined as follows: [2]

Definition 4. *(Laplace Mechanism) Let $\mathcal{Y}$, $\mathcal{Z}$ be two sets, and let $d_{\mathcal{Y}}$ be a metric on $\mathcal{Y} \cup \mathcal{Z}$. Let $\lambda : \mathcal{Z} \to [0, \infty)$ be a scaling function such that $D(y)(z) = \lambda(z)e^{-d_{\mathcal{Y}}(y,z)}$ is a pdf for all $y \in \mathcal{Y}$. Then the mechanism $L : \mathcal{Y} \to \mathcal{P}(\mathcal{Z})$, described by the pdf D, is called a* Laplace *mechanism from $(\mathcal{Y}, d_{\mathcal{Y}})$ to $\mathcal{Z}$.*

We can see that this satisfies $d_{\mathcal{Y}}$-privacy when $d_{\mathcal{Y}}$ is a metric, since

$$\begin{aligned}\frac{D(y)(z)}{D(y')(z)} &= \frac{\lambda(z)e^{-\epsilon d_{\mathcal{Y}}(y,z)}}{\lambda(z)e^{-\epsilon d_{\mathcal{Y}}(y',z)}} \\ &= e^{\epsilon(d_{\mathcal{Y}}(y',z) - d_{\mathcal{Y}}(y,z))} \\ &\leq e^{\epsilon d_{\mathcal{Y}}(y,y')}\end{aligned}$$

when $d_{\mathcal{Y}}$ satisfies the triangle inequality.

In order to gain further understanding about generalised differential privacy, it is helpful to look at an example application. Geo-indistinguishability, introduced next, provides some useful insights into

[2] This is Definition 6 in [12]

how generalised differential privacy can be applied to a particular domain.

4.2 Geo-Indistinguishability

Geo-indistinguishability, developed by Andrés et al. [6], is an application of generalised differential privacy targeting the private release of geo-location points to a data provider. Their motivating example is a user who wishes to receive restaurant recommendations in his local area without an adversary learning his precise location. In this case, the user would expect to send approximate location information in order to receive relevant recommendations, whilst protecting the accuracy with which an adversary can infer his exact location.

Geo-indistinguishability uses the notion of 'l-privacy within radius r' to understand the privacy guarantee provided by generalised differential privacy. More specifically, within a radius r, generalised differential privacy guarantees ϵr-privacy, since for all $x, x' \in \mathcal{X}$ such that $d_{\mathcal{X}}(x, x') \leq r$,

$$d_{\mathcal{P}}(K(x), K(x')) \leq \epsilon d_{\mathcal{X}}(x, x') \leq \epsilon r$$

Geo-indistinguishability can be informally defined as follows [6]:

Definition 5. *A mechanism satisfies ϵ-geo-indistinguishability iff for any radius $r > 0$, the user enjoys ϵr-privacy within r.*

This shows that the level of privacy provided for the user increases as the radius (for privacy) increases. For convenience, we can choose $l = \epsilon r$, thus parametrising the privacy guarantee by l and r. Thus, a user can select a particular l for a *meaningful* radius r and have a privacy guarantee that extends for all distances r. This is a useful insight which permits a practical application of privacy for the user. This notion is referred to as 'l-privacy within radius r'.

In order to formalise geo-indistinguishability using generalised differential privacy, we require a metric over the domain of secrets. For geo-location privacy, the domain of interest is the user's location, specified in geo-location co-ordinates, and a natural metric for this domain is the Euclidean distance metric. Thus, setting $\mathcal{X}$ as the domain of possible locations for the user, and $d_{\mathcal{X}}(x, x')$ as the Euclidean distance between x and x', the authors propose the following definition for geo-indistinguishability:

Definition 6. *(Geo-Indistinguishability) A mechanism K satisfies ϵ-geo-indistinguishability iff for all x, x':*

$$d_{\mathcal{P}}(K(x), K(x')) \leq \epsilon d_{\mathcal{X}}(x, x')$$

where $d_{\mathcal{X}}$ is the Euclidean distance defined on $\mathcal{X}$.

This is similar to the definition of generalised differential privacy, except the ϵ parameter is made explicit, and the domain $\mathcal{X}$ and metric $d_{\mathcal{X}}$ have been defined. The inclusion of the ϵ parameter allows the use of a standard Euclidean metric in the definition; ϵ can then be considered as the 'scaling' factor for the metric.

This definition says that for any output point z, the probability that it came from any point x is similar to the probability that it was produced from another 'close' point x'. In other words, for any co-ordinate output from the mechanism, the user's real location is protected because any location we guess has other 'close' locations which could have produced z with similar probability.

4.2.1 Mechanisms for Geo-Indistinguishability

The Laplace mechanism is a natural mechanism for application to geo-indistinguishability. This mechanism adds noise drawn from a Laplace distribution and is commonly used to apply differential privacy. We saw in Section 4.1.3 that Laplace mechanisms can be used to implement generalised differential privacy using a suitable scaling function. For geo-indistinguishability, the authors use the *polar laplacian*; that is, the application of Laplace noise to the planar co-ordinates (x, y) transformed into polar co-ordinates (r, θ). The authors note that the use of the polar laplacian has been covered in previous work on differential privacy, and this is a standard method for applying Laplace noise in 2 dimensions. They also note that the use of the polar laplacian ensures that the privacy definition is satisfied using the Euclidean distance metric.

4.2.2 Summary

We make the following observations from generalised differential privacy and the example of geo-indistinguishability:

- Geo-indistinguishability demonstrates how to release data privately using generalised differential privacy.
- Generalised differential privacy permits the private release of datasets which consist of a single data point, such as a geo-location co-ordinate. This insight can be extended to author obfuscation, which requires the release of a single obfuscated document.
- The notion of 'l-privacy within radius r' provides a valuable interpretation of the privacy guarantee provided by generalised differential privacy. This insight will be important for understanding authorship privacy.
- The notion of adjacency in differential privacy is replaced by the use of a distance metric for

generalised differential privacy. The Euclidean distance between data points is a natural metric to use in geo-location privacy. For the author obfuscation task, we will look for a corresponding natural metric over the space of authors.

4.3 Author-Indistinguishability

As mentioned earlier in this thesis, the geo-indistinguishability example provides a useful blueprint for understanding how to apply generalised differential privacy. We will see how this can be applied to the author obfuscation problem.

4.3.1 Overview

We first develop some conceptual understanding around how obfuscation will be performed on the input document. Recall the working definition of author obfuscation from Chapter 2:

Definition 1. *(Author Obfuscation) Given a document, obfuscate it in such a way as to prevent identification of its author whilst preserving the topicality of the original text.*

We envisage a mechanism which *adds noise* to the document in some way to produce an output document whose authorship is obfuscated. The addition of noise is a standard method for achieving differential privacy in the literature.

We now consider the utility aspect of the problem; that is, the preservation of topicality. An important observation is that for the author obfuscation problem, the privacy and utility goals, namely authorship protection and topicality preservation, are not directly opposed. This makes the author obfuscation problem distinct from the geo-indistinguishability problem and other problems that we find addressed by differential privacy in the literature. If the privacy and utility goals were orthogonal, then there would be no need for differential privacy; we could simply use randomisation to apply privacy. For author obfuscation, as we noted earlier (see Section 2.1.3), topic-specific words can be identifying for authorship, hence randomisation would destroy (or significantly impact) utility.

In order to align these goals, we can focus on the utility requirement. Recall from our author obfuscation definition (see Section 2.1.3) that the evaluation for utility is performed by a machine learning classifier trained to classify documents by topic. We know that such classifiers typically reduce a document down to a *bag of words* representation, which is a multiset of words. This suggests that the output document from our obfuscation mechanism could be simplified to a bag of words, rather than a semantically sensible document. We also know from Section 2.2.1 that many words which are

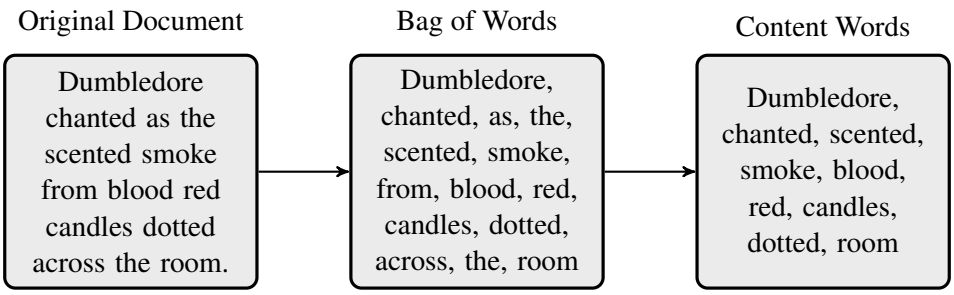


Figure 4.1: Sample transformation of a document to a bag of words with non-content words removed. This transformation discards words unnecessary for topic classification.

useful for identifying authors are non-content words, such as prepositions and pronouns. This suggests that we can further simplify our output document by removing non-content words which do not contribute to topicality, but which can be used to identify authorship. These steps will be performed as pre-processing steps prior to applying an obfuscation mechanism, as shown in Figure 4.1.

These pre-processing steps not only simplify the language component of the problem, but also align the privacy and utility goals. That is, all of the words in the pre-processed document can be considered identifying for both topicality and authorship. Note that we expect the pre-removal of some non-content words to impact on the ability of an adversary (armed with an author identification tool) to guess document authorship. However, note also that we do not consider this a privacy step, as no randomisation has taken place. In the experiments in the next chapter we go into further detail on how the documents will be transformed prior to obfuscation.

4.3.2 Comparison with Geo-Indistinguishability

We now make the connection between geo-location privacy and author obfuscation. Recall that geo-indistinguishability defines privacy for geo-location co-ordinates; that is, points in a 2-dimensional space. Privacy is protected for the release of a data point x corresponding to the user's location. This data point is obfuscated such that the output seen by an adversary could have come from any 'close' point with (almost) equal likelihood.

For author obfuscation then, we can consider a notion of author-indistinguishability that would entail protecting the release of a document x which is obfuscated such that the output document seen by an adversary could have come from any 'close' author with (almost) equal likelihood. We envisage adding noise to the document, in some way, so as to protect the author's identity. This has conceptual similarities with geo-indistinguishability, however we note some key differences.

Firstly, notice that for geo-indistinguishability, the domain of location co-ordinates is common to both the obfuscation mechanism (which takes points as inputs and returns points as outputs) and the

distance metric (since privacy is with respect to the distance between points). However for author-indistinguishability we anticipate a mechanism that operates on the domain of *documents* (that is, taking documents as inputs and returning obfuscated documents as outputs) but a distance metric that operates on the domain of authors (since privacy is with respect to the distance between authors). This mismatch between documents and authors needs to be resolved in order to make sense of a definition of author-indistinguishability based on geo-indistinguishability.

Secondly, a key concept in geo-indistinguishability is the notion of 'l-privacy within radius r'. This concept is aided by the representation of geo-location co-ordinates as points in 2-dimensional space. We would like a corresponding geometric understanding of authors as points in n-dimensional space.

Finally, generalised differential privacy stipulates the use of a metric, which we require over the space of authors. This is particularly relevant for the application Laplace mechanisms.

4.3.3 Stylometry Revisited

We will now revisit some concepts from stylometry discussed in Chapter 2 (see Section 2.2.1) and introduce some key ideas in order to address the above requirements.

Authors as Documents

A key notion used in the stylometry literature for author identification is that *authors are identified by the documents they write*. In other words, an author can be considered to be a collection of documents. This notion is used in author identification methods in order to identify documents by particular authors; in this way, unknown documents can simply be compared with other documents by known authors to determine if there is a match for authorship. There are a number of ways in which this notion is realised in the literature. Some author verifiers represent authors as a single document, either by concatenating documents by the same author together, or by computing the *document centroid* of the documents by that author. Other methods utilise sets of documents as identifiers for an author [35]. The Koppel Method [3] selected for this task represents each author as a single document and uses the character n-gram for that author as a 'fingerprint' for determining whether an unknown document is by the same author. [4] The character n-gram representation provides a way to represent documents in such a way as to permit authorship identification. In this way, it provides the link between authors and documents which resolves the document-author mismatch identified above.

[3] Recall from Section 2.2.2 that the Koppel Method is the method by Koppel et al. [36].

[4] Although the Koppel Method uses character 4-grams, we will continue to use the more general term n-gram to describe this representation, leaving the 4-gram specifics to our experimental implementation.

Vector Representations

Recall from Section 2.2.1 that it is standard practice in both stylometry and in machine learning to represent documents by *feature vectors*. For the Koppel Method, these feature vectors are character n-grams, and are used to represent authors (as documents). Typically the n-gram vectors for a document set will contain tens of thousands of n-grams (dimensions); Koppel et al. [36] reduce these to the 100,000 most frequently occurring n-grams across the document set. The n-gram vectors can either be integer-valued (representing the frequency counts of particular n-grams) or real-valued (representing the *tf-idf* counts for n-grams). [5] These n-gram vectors are treated as vectors in the 'space' of authors when computing similarity measures for documents. For example, the Koppel Method measures similarity using the *cosine similarity* between the document-author vectors, where the cosine similarity is simply the normalised dot product of the vectors. Thus we can treat authors as points in high-dimensional space using their n-gram vector representations.

Metrics over Authors

Author verification methods rely on distance measures in order to determine the 'nearest neighbour' to a document. As mentioned above, the Koppel Method uses cosine similarity to assess the closeness of a document to an author. An extension of this method by Koppel & Winter [35] shows that the *minmax similarity* is a more effective measure for similarity. Whilst the minmax similarity is not a metric, its complement, known variously as the Ruzicka metric [33], or the generalised Jaccard distance [37], *is* a metric. [6] Moreover, this metric has been used successfully with Koppel & Winter's character 4-gram method for authorship identification [33]. This metric is formally defined as follows:

Definition 7. *(Ruzicka Metric) Let $\mathcal{X}$ be the set of all documents and let x, y be documents in $\mathcal{X}$. Let n be the number of features in the chosen document representation for $\mathcal{X}$, and let $\vec{x} =< x_1, x_2, ..., x_n >$, $\vec{y} =< y_1, y_2, ..., y_n >$ be vector representations of the documents x, y respectively. Then the Ruzicka metric on $\mathcal{X}$ is defined as:*

$$Ruzicka(\vec{x}, \vec{y}) = 1 - \left(\frac{\sum_{i=1}^{n} min(x_i, y_i)}{\sum_{i=1}^{n} max(x_i, y_i)} \right)$$

Note that the Ruzicka metric can be used with either frequency counts or tf-idf values in the vector representation.

[5] Tf-idf, or term-frequency inverse-document-frequency, is a common measure used in NLP for weighting the relative importance of features.

[6] The metric properties of the Ruzicka metric can be derived from the proofs presented in [37].

4.3.4 Summary

Using the Koppel Method, which involves representation of documents using character n-gram vectors, we can resolve the issues raised in Section 4.3.2. The notion that authors are the documents they write allows us to resolve the mismatch of domains for author-indistinguishability. Thus the inputs and outputs of the obfuscation mechanism can be treated as 'authors' by using a character n-gram representation for documents. Also, this representation can be treated as a high-dimensional vector space using either integer-valued or real-valued vectors. And finally, the Ruzicka metric can be used to measure the distance between authors in the n-gram vector space. Thus the notion of 'l-privacy within radius r' can be meaningfully interpreted in terms of the 'author space' of n-gram vectors.

4.3.5 Author-Indistinguishability Defined

We now have a better understanding of the connection between geo-location and author obfuscation. Authors have a vector representation via character n-grams, and a corresponding metric in the Ruzicka metric which can be used to measure distance between authors in the character n-gram space. We are now ready to define author-indistinguishability.

Let $\mathcal{X}$ be the set of possible documents, let $\mathcal{Z}$ be the set of possible output documents, and let $\mathcal{X}_A$ be the set of character n-gram representations of documents in $\mathcal{X}$. Let $\mathcal{P}(\mathcal{Z})$ be the set of probability measures over $\mathcal{Z}$. Let $K : \mathcal{X}_A \to \mathcal{P}(\mathcal{Z})$ be a probabilistic function known as a *mechanism*. Let $A : \mathcal{X} \to \mathcal{X}_A$ be a mapping from the (canonical) representation $\mathcal{X}$ to the character n-gram (vector) representation $\mathcal{X}_A$. Let d_A be the Ruzicka metric defined on $\mathcal{X}_A$.

Definition 8. *(Author-Indistinguishability) A mechanism $K : \mathcal{X}_A \to \mathcal{P}(\mathcal{Z})$ satisfies ϵ-author-indistinguishability iff for all x_A, $x'_A \in \mathcal{X}_A$:*

$$d_{\mathcal{P}}(K(x_A), K(x'_A)) \leq \epsilon d_A(x_A, x'_A)$$

This definition says that authors which are 'close' in the character n-gram space should have similar outputs from the mechanism K. Alternatively, given an output z, it should not be possible to determine whether it came from author x or author x', where x and x' are 'close' in author space. Note that although this definition treats the inputs and outputs of K as authors, they can equally be thought of as documents.

The notion of 'l-privacy within radius r' can now be understood for author-indistinguishability. Given a document to obfuscate, we can represent the document as an author using a character n-gram vector. We can then choose other authors (that is, documents by other authors) against which we want the document to be indistinguishable. By converting those documents into their character n-gram vectors,

we choose r to encompass these documents with respect to the Ruzicka metric. Then any choice for ϵ gives ϵr-privacy within this radius r, or alternative l-privacy within radius r for $l = \epsilon r$.

We might ask whether documents should be obfuscated with respect to *any* author, and not just close ones. However, recall that there is also a utility measure to consider, namely preservation of topicality. Obfuscation with respect to any author would disregard topicality and thus destroy the utility of the output document.

4.3.6 Mechanisms for Author-Indistinguishability

As for geo-indistinguishability, we consider mechanisms for adding noise to the document. The natural mechanism to consider is the Laplace mechanism, as is also used in the geo-indistinguishability example. However, unlike geo-indistinguishability, this problem involves higher dimensional space; as mentioned earlier, the Koppel Method typically employs 100,000-dimensional n-gram vectors.

For simplicity, we consider a noise mechanism which involves adding noise to each dimension of the character n-gram vector. This is a technique which has been applied to vectors in the literature using the Laplace mechanism [24]. Let's first define this problem more specifically.

Preliminaries

Let $\mathcal{W}$ be a set of words (known as a vocabulary) and $d \in \mathcal{W}^N$ be a document to obfuscate (represented as a sequence of words). Let d_i denote the ith word in the document d. Let $x \in \mathbb{Z}^k$ be the n-gram vector for d and let $\mu \in \mathbb{Z}^k$ represent some amount of noise to add to x in order to satisfy differential privacy. That is, $K(x) = x + \mu$. Let $M : \mathcal{W} \to \mathbb{Z}^k$ be a mapping from each word to its n-gram vector and let $Sim : \mathcal{W} \to \mathbb{P}(\mathcal{W})$ be a mapping from each word to a subset of 'similar' words (also containing the original word).

Definition 9. *(WordSum Problem) Given the mappings M and Sim, an input document d and a noise vector μ, output a new document d' such that*

$$d'_i \in Sim(d_i) \quad \forall i \in N \qquad \textit{and} \qquad \sum_{i=1}^{N} M(d'_i) = \sum_{i=1}^{N} M(d_i) + \mu$$

In other words, the problem is to map an input document to a (noisy) output document by modifying the n-grams in the author vector. The amount of noise to be added should be calculated using the n-gram vector and the words chosen so that their n-gram representations sum to the required (noisy) n-gram vector. However, it turns out that this problem is NP-hard.

Theorem 1. *The WordSum problem is NP-hard.*

Proof. Consider the Subset Sum problem (which is NP-complete): given $n \in \mathbb{Z}$ and a set $S = \{z_1, z_2, ..., z_N\}$ where $z_i \in \mathbb{Z}$, find $S' \subseteq S$ such that $\sum_{z_i \in S'} z_i = n$. We show how to reduce this problem to the WordSum problem.

Assume that the WordSum problem can be performed in polynomial time wrt N and let $k = 1$. Consider the Subset Sum problem defined above. Given $S = \{z_1, z_2, ..., z_N\}$, choose any (unique) words $d = \{w_1, w_2, ..., w_N\}$ s.t. $w_i \in \mathcal{W}$ and define $M : \mathcal{W} \to \mathbb{Z}$ by $M(w_i) = z_i$ for $1 \leq i \leq N$. Choose any $\perp \in \mathcal{W}$ s.t. $\perp \notin d$ and define $M(\perp) = 0$. Now, define $Sim : \mathcal{W} \to \mathbb{P}(\mathcal{W})$ by $Sim(w_i) = \{w_i, \perp\}$ and choose $\mu = n - \sum_{i=1}^{N} M(w_i)$.

The above mappings can be performed in linear time (wrt N).

Using M, Sim, d and μ as inputs to the WordSum problem, the output will be a document d' satisfying $d'_i = w_i$ or $\perp$ for all $1 \leq i \leq N$ and $\sum_{i=1}^{N} M(d'_i) = n$. We then have that $S' = \{M(d'_i) \mid d'_i \neq \perp\}$, where S' is the solution to the Subset Sum problem This can be performed in linear time (wrt N).

Therefore, if the WordSum problem can be solved in polynomial time, so can the Subset Sum problem. Hence the WordSum problem is NP-Hard.

□

Given the dimensions of the word and n-grams spaces (approximately 3 million words in our synonym set and 100,000 dimensional n-gram vectors) this problem seems intractable. In addition, we have no other mechanisms for applying Laplace noise to draw on for this problem. Instead, let's re-examine the author-indistinguishability definition to see if it can be reframed.

4.3.7 Author Indistinguishability Re-Examined

We have seen that adding noise to the character n-gram vectors is computationally hard, because the noisy documents cannot be recovered. Let's re-consider author obfuscation in the light of document representations. Recall that authors are considered to be the documents that they write, and the difference between these two concepts lies in their representations. We ask the question: can obfuscation in document space result in obfuscation in author space? That is, can we define author indistinguishability in terms of *document indistinguishability*?

We first need to understand what document indistinguishability means. Intuitively, document-indistinguishability says that the outputs from a privacy mechanism K on input documents x and x' depends on their distance $d_{\mathcal{X}}(x, x')$. We should choose a distance metric that preserves semantic similarity in document space in order to preserve the topicality of the output document. Document-indistinguishability, then, promises that semantically similar documents produce similar distributions.

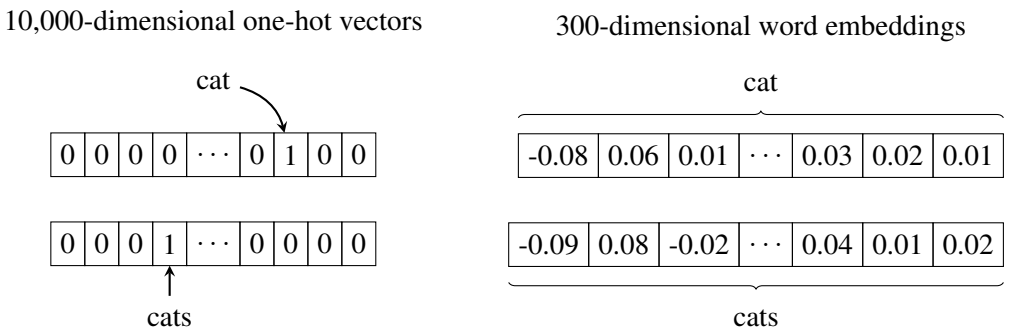


Figure 4.2: One-hot vector representation of 'cat' and 'cats' versus word embedding representations. Word embeddings encode semantic relationships, hence similar words have higher cosine similarity. One-hot vectors are all orthogonal and do not encode relationships between words.

The application of generalised differential privacy to this problem requires firstly an appropriate representation which permits documents to be represented as points in space, and secondly a metric defined over that representation which measures semantic similarity between documents.

Word Embeddings

The NLP community has long been interested in document representations, which are useful in many applications involving document searching and processing. Document representations are typically word-based, as opposed to the character n-gram based representations that are used in author identification. An important new direction in NLP is the move towards compact vector representations of words known as *word embeddings*. Word embeddings replace traditional vector representations of words, which use sparse 'one-hot' vector representations to encode words. In contrast, word embeddings encode words as lower dimensional vectors which are learned using neural networks trained over very large datasets. One reason for the rise in popularity of these word representations is the interesting semantic properties that they entail. This is because word embeddings are learned using the notion that semantically similar words are found in similar contexts, and as a result, semantically similar words have similar representations. This feature is absent from one-hot representations, in which no relationships between words are captured. For example, the words 'cat' and 'cats' have arbitrary representations in one-hot encodings, but their word embedding vectors are similar due to the similarity in their contextual use. This example is shown in Figure 4.2.

One word embedding implementation commonly used in the literature is Word2Vec [47]. A key contribution of Word2Vec is that it aims to preserve semantic relationships between word embeddings so that vector operations can be meaningfully used on word vectors. For example, in Word2Vec, vector('king') - vector('man') + vector('woman') is closest to vector('queen'). Similarly, vector('Paris') - vector('France') + vector('Italy') is closest to vector('Rome'). Word2Vec can be used to find the most similar words to a given target word. Table 4.1 shows an example of semantically similar words

encoded in Word2Vec. Note that these semantically similar words are found using the cosine similarity score (related to angular distance) between the target vector and each word in Word2Vec. An additional advantage of Word2Vec is that it comes with pre-trained vectors which can be used 'off the shelf'. This is particularly helpful as learning word embeddings using neural networks requires enormous quantities of data.

Target	**car**	**run**	**understand**	**because**	**town**
Closest words	vehicle	runs	comprehend	but	village
	cars	running	explain	so	hamlet
	SUV	drive	know	Because	towns
	minivan	scamper	realize	anyway	city

Table 4.1: Sample words and their closest word embeddings, showing that semantic relationships are captured by similar word embedding vectors.

Documents using Word2Vec representations are typically encoded as a bag of word embeddings, and are represented as vectors with each co-ordinate holding the word count for a particular word in the vocabulary. In this way, documents can be considered as points in an n-dimensional space. The use of word embeddings permits particular *metrics* over the document space, which provide measures of semantic distance by exploiting the semantic nature of the word embedding vectors.

Document Metrics

A number of document similarity measures are commonly used in NLP tasks involving semantic similarity, and there is no clear consensus on any particular measure in the literature. However, our decision to use word embeddings strongly suggests the use of a metric which incorporates word embedding features. The Word Mover's Distance (WMD), proposed by Kusner et al. [38], is a metric based on the Earth Mover's Distance, and is designed for use with word embeddings. The WMD between two documents d and d' is defined as the minimum distance required to move all the words (represented as word embeddings) from d to d'. An example is shown in Figure 4.3. The distance between word embeddings is defined here as the Euclidean distance, although this metric is not commonly used to represent similarity in Word2Vec. However, WMD has shown state-of-the-art performance compared with other document similarity measures, and is a natural choice for use with Word2Vec. Its main drawback is computation time; WMD is based on the Earth Mover's Distance [55], and is framed as an optimisation problem using linear constraints. The solution is slow to compute with average time complexity $O(p^3 \log p)$, however the authors propose optimisations which can reduce the size of the problem.

4.4 Document-Indistinguishability

We are now ready to define document-indistinguishability.

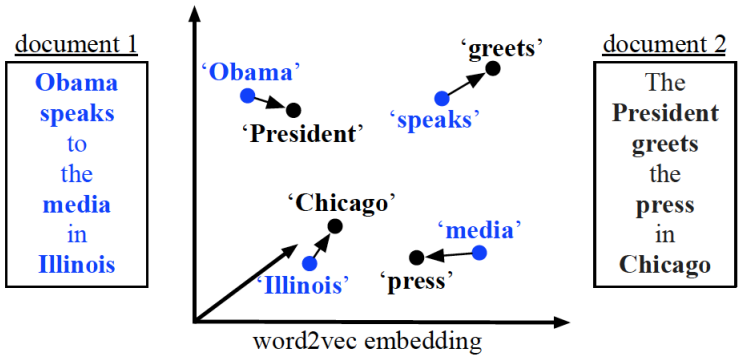


Figure 4.3: Example of Word Mover's Distance taken from [38]. The documents (ignoring stopwords) are transformed into word embeddings and compared in Word2Vec space. The distance between the two documents is the minimum Euclidean distance that words in document 1 need to travel in order to match the words in document 2.

Let $\mathcal{X}_D$ be the space of documents represented using word embeddings and d_w be the Word Mover's Distance defined on $\mathcal{X}_D$. Then we define document-indistinguishability as follows:

Definition 10. *(Document-Indistinguishability) A mechanism $K : \mathcal{X}_D \to \mathcal{P}(\mathcal{Z})$ satisfies ϵ-document-indistinguishability iff for all x_D, $x'_D \in \mathcal{X}_D$:*

$$d_{\mathcal{P}}(K(x_D), K(x'_D)) \leq \epsilon d_w(x_D, x'_D)$$

This definition says that documents which are close together with respect to the Word Mover's Distance produce output distributions that are close together. Alternatively, any output z is (almost) just as likely to have come from a particular document x as it is to have come from another semantically similar document x'.

However, we need to relate this definition to the the original problem: author obfuscation. We can suggest how this might be done using the notion of 'l-privacy within radius r'. Given a document x, we would like to choose a set of documents which are to be 'indistinguishable' with respect to the mechanism K. We want to choose documents that are by more than one author so that we can also achieve author-indistinguishability. We can then choose r using the distance measure in the document representation, knowing that documents within radius r, and hence *authors* within radius r are indistinguishable. We present this below as a hypothesis.

Hypothesis 1. *ϵ_1-Document-indistinguishability $\Rightarrow$ ϵ_2-author-indistinguishability for some $\epsilon_1 < \epsilon_2$.*

The intuition, then, is that some (possibly large) amount of noise could be added to documents using

a word embedding representation, and this would provide some weaker privacy guarantee for author-indistinguishability. We will examine this hypothesis experimentally in the next chapter.

4.4.1 Mechanisms for Document-Indistinguishability

As for geo-indistinguishability, we will consider using a Laplace mechanism for this problem, recalling that any Laplace mechanism satisfies generalised differential privacy (Section 4.1.3).

Firstly we considered how document-indistinguishability could be applied by adding noise to individual words in the document. The geo-indistinguishability example used the polar laplacian in 2 dimensions, however we require Laplacian noise for higher dimensions, considering the document as a vector of words. We could not find guidelines in the literature on how to add Laplace noise for higher dimensions, and the conversion to polar co-ordinates appears to be non-trivial.

As a second option, we considered adding Laplace noise to each dimension of the vector, as demonstrated by Dwork [24]. However, we note that the addition of noise separately to each component of a vector is only valid using the Manhattan distance metric on the space of elements. In our definition we use the Word Mover's Distance, and it is not clear how we would be able to add noise using this metric. In fact there is no pre-existing work on Laplace noise for non-Euclidean metrics, which would prevent our use of alternative similarity measures commonly used in text document processing.

Although we were unable to construct a Laplacian mechanism which satisfies our definition of document-indistinguishability, we will still explore the use of Laplace noise experimentally in the next chapter. This will be used to determine the feasibility of using Word2Vec as a mechanism in the future.

4.4.2 Summary

Generalised differential privacy and the application of geo-indistinguishability demonstrate a new way of approaching differentially private data publishing through the private release of individual data points. This insight motivated our exploration of author obfuscation as a similar problem requiring the private release of individual documents. We considered the definition of author-indistinguishability in the light of geo-indistinguishability, however we found that implementing a noise-based mechanism using character n-gram vectors is NP-hard. We then turned to the question of document-indistinguishability, and asked whether this could be applied to the author obfuscation problem. We did not prove this connection, but we will explore this experimentally in the next chapter. Finally, we considered appropriate mechanisms using Laplace noise, however we found that the Laplace mechanisms described in the literature require the use of a Euclidean metric. This represents a gap in the literature that impacts our use of Laplace mechanisms in the text document domain, which makes heavy use of non-Euclidean measures for similarity.

5 EXPERIMENTAL RESULTS

This chapter covers the experimental setup and methodology used to evaluate aspects of the theoretical framework proposed in the previous chapter. The purpose of these experiments is to explore the feasibility of applying generalised differential privacy to the author obfuscation problem. In particular, we are interested in evaluating whether Word2Vec can be used to generate noisy documents in such a way as to preserve some semantics of the original text, through preservation of its topicality. We also investigate the hypothesis from the last chapter, in which we propose that the application of privacy to documents implies privacy at the level of authorship.

5.1 Overview

In the previous chapter we hypothesised that applying obfuscation to a document (so as to provide document-indistinguishability) also results in obfuscation at the level of authorship, perhaps for a reduced degree of privacy. We also saw that reducing a document to a bag of content words aligns the privacy and utility goals by discarding those parts of the document which are irrelevant to utility but which could compromise privacy. In this chapter, we aim to explore the feasibility of implementing author privacy using tools from natural language processing. We would like to answer the following questions:

1. Is there a relationship between the document space metric (Word Mover's Distance) and the author space metric (Ruzicka Metric)? This would allow us to understand privacy in author space in terms of obfuscation performed in document space (using document-indistinguishability).

2. Can Word2Vec be used to generate obfuscated documents using the semantic similarity properties of its word embedding vectors? This would allow us to make use of existing NLP technology and head towards a fully automated obfuscation mechanism.

In these experiments we implement a simple obfuscation mechanism which adds noise to individual words in the document and fetches the generated 'noisy' words from Word2Vec. We then evaluate the obfuscation mechanism using the metrics outlined earlier in this thesis (Section 2.1.3), namely:

- *Safeness* - that is, whether the original author can be identified from the obfuscated text. This will be evaluated using a single state-of-the-art author verification tool rather than a suite of automated verifiers.

- *Soundness* - that is, whether the obfuscated text preserves the topicality of the original text. This will be evaluated using a machine learning classifier trained on documents labelled by topic.

5.2 Datasets

We require datasets which are labelled with topics as well as authors in order to be able to measure topicality and authorship. We used two datasets for these experiments. One dataset is commonly used in the NLP literature, the other one we constructed as we were unable to find a second appropriate dataset labelled with both topics and authors. The datasets we used are described below:

1. The Reuters RCV1 dataset is a standard dataset used in language processing tasks, and consists of over 800,000 Reuters news articles separated into various topics [39]. Although not originally constructed for author attribution work, it has been used previously in this domain by making use of the *<byline>* tags inside articles which designate article authors [58]. The dataset was chosen because it contains documents of reasonable length, which is required for successful author identification. In addition, this dataset is similar to the dataset on which the Word2Vec vectors used in this experiment were trained on, and thus we would expect high quality outputs when using Word2Vec with this data.

2. Fan fiction from https://www.fanfiction.net was collected and used to construct a dataset consisting of stories collected over the 5 most popular book-based topics. Fan fiction has been used previously in PAN author attribution tasks, and is suitable for this task because of the content length of the texts and the diversity of authorship styles present in these texts, as stylistic writing qualities are important in this domain. The dataset is also similar to the `blogger.com` dataset used in author attribution evaluation by Koppel & Winter [35], however that dataset does not contain topic labels as required for evaluation of our mechanism.

For each dataset, we required separate training and test sets, as is standard practice in machine learning. Training sets are used to train the classifier and produce a model, which is then run on unseen (test) data in order to compute a result. Although it is standard practice to also use a development set for parameter setting, we did not require this as we chose not to vary the parameters of the classifiers used in these experiments. Note that the separation of training and test sets was required for the topic classifier, which is machine learning based, but not the authorship attribution approach, which uses the 'training' data as a known set against which to test unknown documents. However, the terms 'training' and 'test' will be used throughout this chapter for consistency.

For the Reuters dataset, a subset of the RCV1 dataset was chosen. The training dataset consisted of 2000 documents evenly spread across 5 topics, with 20 authors per topic and 20 documents per author. The test dataset was an even spread of 500 documents using the same 5 topics and 20 authors as for the training dataset.

For the Fan fiction dataset, the training and test sets were constructed following the setup used by Koppel et al. [36] In particular, for each author we chose 2000 words of writing for the training dataset, and a different 500 words for the test dataset. Only authors with at least 2500 words of writing were selected for this dataset. This resulted in a training dataset of 102 documents from 100 authors spread across 5 topics and a test set of 102 documents from the same 100 authors spread across 5 topics. [1]

A summary of these datasets is shown in Table 5.1.

Dataset	Topic	# Training	# Test	# Authors
Reuters	C11 (Strategy/plans)	400	100	20
	C12 (Legal/judicial)	400	100	20
	C13 (Regulation/policy)	400	100	20
	C21 (Production/services)	400	100	20
	C24 (Capacity/facilities)	400	100	20
Total		2000	500	100
Fan fiction	Harry Potter	22	22	22
	Hunger Games	22	22	21
	Lord of the Rings	15	15	15
	Percy Jackson and the Olympians	14	14	13
	Twilight	29	29	29
Total		102	102	100

Table 5.1: Summary of training and test dataset splits by topic and authors.

Note that although these datasets are small for topic classification tasks, they represent large datasets with respect to the author attribution literature. In comparison, the PAN 2016 only contained documents across 14 different authors, and Koppel et al. [36] note that the much work in the author attribution literature uses datasets consisting of only a handful of authors. Given that the primary purpose of this task is author obfuscation, the small dataset size is acceptable with respect to other experimental work in the literature.

The documents were formatted according to the requirements of the PAN 2012 author attribution task. This was to allow author attribution software from PAN to be tested on the data, in particular the code base for the author attribution method by Koppel et al. [36] Note that the PAN datasets could not be used for these experiments as they lack topic labels for the data.

Note that publishing restrictions on the Reuters RCV1 dataset require that no complete data snippets from RCV1 can be included in this thesis, although results from analysis can be included. Document snippets in this section are drawn from the Fan fiction dataset, however individual words and obfuscations of words from the Reuters dataset have been included where appropriate.

[1] The extra 2 documents are due to our need to collect multiple stories by one author in order to reach 2500 words.

5.3 Methodology

After data collection, we performed some post-processing to remove words unnecessary for topic classification, as discussed in the previous chapter (Section 4.3.1). We first removed non-content words from the documents using a 'stopword' list. Stopwords are words with minimal lexical value, such as prepositions and pronouns, and are commonly removed for NLP tasks such as topic classification. The stopword list used for this step is shown in Appendix A, and is the standard list supplied with the Python Scikit-Learn toolkit used in these experiments. [2] The resulting dataset is referred to as the Content-Words dataset.

We then created a second modified dataset by removing the least useful words for topic classification from the Content-Words documents. We identified these words using *feature selection*, which is commonly used in NLP tasks to improve the performance of classifiers [65]. A feature selector identifies the features which are most discriminative for the classification task. Discarding less useful features often has minimal impact on accuracy but significantly improves performance by reducing the dimensionality of the feature set. We chose to use the *chi-squared feature selector* since it has been identified as high-performing and has the advantage of being independent of any particular machine learning classifier [65]. Chi-squared feature selection makes use of the chi-squared (χ^2) statistic to score the dependence of each word on each class; the highest scoring words are those with the most discriminative power, and thus are most useful for classification. Whilst chi-squared is an older feature selection method, it is still considered state-of-the-art and is one of the preferred methods for text classification tasks in the literature [15]. The chi-squared feature selector was used to identify the best (most discriminating) 'n' words in the document set, which was then modified so that each document contained only those words in common with the 'n' best words. The resulting datasets are referred to as BOW-n datasets.

The datasets created from the Reuters and Fan fiction datasets are summarised below:

1. *Raw* - Documents containing raw text, stripped of html and special characters. These were used for baseline evaluation.

2. *Content-Words* - Documents stripped of non-content words. This was performed using a standard stopword list, which is shown in Appendix A.

3. *BOW-n* - Documents converted into bags of topic words using the n best features for topic classification. We used values of n = 1000, 500, 200 and 50 for evaluation. For example, documents in the BOW-1000 dataset contain only words which are in the 1000 most discriminative feature words for topics in the dataset. Note that as n decreases, then we expect the average length of

[2] Scikit-Learn is a standard Python toolkit for machine learning. http://scikit-learn.org/

documents to decrease, corresponding to a smaller set of matching words for each document.

This post-processing was performed on both the training and test datasets. Some sample document snippets from each of these datasets are shown in Table 5.2.

Raw	Content-Words	BOW
It was nearing two in the morning, and ... the inside of Weasley's Wizard Wheezes was just as lively as ever.	nearing morning inside weasley wizard wheezes just lively	morning weasley wizard
The words, "Bring forth the ring, Frodo," caused the Hobbit startle and swallow thickly.	words bring forth ring frodo caused hobbit startle swallow thickly	ring frodo

Table 5.2: Sample documents from the Raw, Content-Words and Bag of Topic Words (BOW) datasets. Column 1 shows a raw text document snippet with all html tags stripped. Column 2 shows the same document snippet with non-content words removed. Column 3 shows the document snippet with only topic-specific words remaining.

Redaction of Proper Nouns

We noted that the most discriminating words across the Fan fiction dataset corresponded to proper nouns such as names of characters. These words did not dominate the Reuters dataset as noticeably. We assume this is because the character names used in Fan fiction are exclusive to particular books, whereas proper nouns such as country names or names of businesses can occur across topics in the Reuters dataset. For this reason, we performed a proper noun redaction across all of the Fan fiction datasets prior to generating bag of words documents. We did this to isolate the effect on topicality of the obfuscation, which would otherwise be dominated by unobfuscatable proper nouns. In particular, any arbitrarily chosen alternative name (such as John for Harry) will not preserve the topicality of the document, as required for soundness.

5.3.1 Obfuscation Mechanism

Next, we defined an obfuscation mechanism. Recall from Section 4.4 that document-indistinguishability applies obfuscation at the word level using a word embedding representation and the Word Mover's Distance metric. We chose to use 300-dimensional pre-trained Word2Vec word embeddings, [3] which are widely used in machine learning tasks. These word embeddings were trained on a 3 billion word corpus of news articles and contain a total of 3 million words. The training of word embeddings requires enormous resources, so the use of pre-trained embeddings is commonplace. [4]

We considered two possible mechanisms for obfuscation of words in each document. Note that we do not claim privacy guarantees for either of these mechanisms, but we investigated them as they

[3] Available from https://code.google.com/archive/p/word2vec/

[4] Training embeddings from data in the same domain as a task can sometimes result in better classification performance, but we do not explore that here.

are simple adaptations of Laplace mechanisms, and thus represent the sorts of mechanisms that we envisage being used in future work with privacy guarantees.

The first mechanism applies noise to individual words by calculating a noisy angle θ, computing its cosine and querying Word2Vec for the closest word at this cosine distance from the original word. We selected the noisy angle from a Laplace distribution centred at 0 and truncated to $\frac{\pi}{2}$. This means that values of θ closer to 0 are more likely than larger values for θ; this in turn implies that words with higher cosine similarity, and hence higher semantic similarity, are more likely to be chosen. We consider this criteria for selecting noisy words the better of the two mechanisms, as it more closely aligns with the similarity measure used in Word2Vec, and thus noisy words selected are more likely to be semantically related. This mechanism is illustrated in Algorithm 1.

Algorithm 1 Obfuscation Mechanism 1

Require: radius r, epsilon ϵ, word embeddings $word2vec$, documents d
  for doc in d **do**
    words = list words in doc
    for w in words **do**
      noisy_theta = Lap(r/ϵ) truncated to $[\frac{-\pi}{2}, \frac{\pi}{2}]$
      sim = cosine(noisy_theta)
      noisy_word = lookup closest word in $word2vec$ at distance sim from w
      add noisy_word to noisy_doc
    end for
    add noisy_doc to obfuscated dataset
  end for
  return obfuscated dataset

The second mechanism we considered treats each word as a 300-dimensional vector and applies Laplace noise to each component of the vector. We then query Word2Vec for the closest vector to this noisy vector, and retrieve the word corresponding to that vector. We considered this mechanism because it more closely represents a Laplace mechanism applied to vectors in Euclidean space. However, Word2Vec similarity is not measured using Euclidean distance, hence this noise-adding mechanism is not ideal for our scenario (and, in particular, we note that the results for Laplace mechanisms do not apply here). This mechanism is illustrated in Algorithm 2.

Note: Although mechanism 1 represented a better choice for the generation of noisy words, we found this mechanism too computationally expensive to implement using the Word2Vec API chosen as it required searching the space of words for the closest word with the generated cosine similarity. Mechanism 2 was thus the only mechanism implemented for all of the experimental work below. We simply refer to this as the obfuscation mechanism for the remainder of this chapter.

Note also that we used the gensim API [5] for our Word2Vec implementation.

[5] https://radimrehurek.com/gensim/models/word2vec.html

Algorithm 2 Obfuscation Mechanism 2

Require: radius r, epsilon ϵ, word embeddings $word2vec$, documents d
  for doc in d **do**
    words = list words in doc
    for w in words **do**
      vec = vector for word w in $word2vec$
      noisy_vec = add noise from Lap(r/ϵ) to each dimension in vec
      noisy_word = lookup closest word to noisy_vec in $word2vec$
      add noisy_word to noisy_doc
    end for
    add noisy_doc to obfuscated dataset
  end for
  return obfuscated dataset

The obfuscation mechanism was run over the test data for the Reuters and the Fan fiction datasets and stored for evaluation. We found that Word2Vec was slow in computing the closest words given a vector of interest, so we were unable to run the mechanism 'on the fly' to generate obfuscated data. We would have preferred to be able to run the mechanism multiple times for each dataset in order to present an average result for each dataset, however this was not possible due to time constraints. As a consequence, the results are reported for only a single run of obfuscation over each dataset. Due to the inherent randomness in the obfuscation mechanism, we expect some results may be inconsistent.

Parameters

The obfuscation mechanism is parametrised by the radius r and ϵ parameters; these were combined into a **scale** parameter using scale = $\frac{r}{\epsilon}$ which was passed to the mechanism and used to parametrise the Laplace distribution. Scales of 0.1, 0.2 and 0.5 were chosen after experimentation revealed that at scale 0.1 most of the words remain unchanged, whilst for scale 0.5 large changes were noticed in the text. The scale parameter will be passed to the obfuscation mechanism used in these experiments, and results are reported with respect to this parameter. Note that higher scales correspond to more noise and hence more obfuscation.

5.3.2 Evaluation

We next selected methods for evaluating the output from the obfuscation mechanism for *safeness* and *soundness*.

The author identification method chosen for evaluation of safeness was an implementation of the approach by Koppel et al. [36] which we described earlier (Section 2.2.2) and refer to as the Koppel Method. This implementation is available on the PAN website [6] and is one of the verifiers used in the

[6] https://github.com/pan-webis-de/koppel11

suite of verification software for evaluating the PAN author obfuscation task.

The Koppel Method creates a feature set consisting of the N most frequent character 4-grams used in the dataset. Then, k random subsets of n features from the feature set are selected and used to test authorship. The intuition behind using random subsets of features is that a document should match its author regardless of the feature set chosen, and therefore author identification must be robust to variations in the underlying feature set. Thus, the most likely author can be selected as the most commonly attributed author across the k feature sets. If no author matches at least t times (for some threshold value t), the document is assigned 'unknown author'.

The original algorithm uses the cosine similarity to compute the distance between an unknown text and a known text. [7] We modified this algorithm to use the Ruzicka metric, which is the complement of the minmax similarity; the minmax similarity was shown to perform better at this task [35], however this does not have the metric properties we require in a distance measure. [8]

The modified algorithm for our Koppel Method implementation is shown in Algorithm 3.

Algorithm 3 Koppel Method using the Ruzicka Metric

Require: list of known documents dl for range of candidates C, unknown document u
  for 1 to k **do**
    randomly choose n features from the full feature set
    find the closest document d to u using the Ruzicka metric
    record c, the author of d
  end for
  for each candidate c in C **do**
    score(c) = proportion of times c was the closest author
  end for
  return author with max score

The Koppel Method codebase has a number of parameters to set. We used $k = 100$ and $n = N/2$, as suggested in [35]. We ignore the threshold t, assuming the closed-world scenario (that is, where we always guess an author). We also restricted the maximum number of features N to 20,000 for efficiency, noting that reduced feature sizes decrease the accuracy of the algorithm [35].

The machine learning classifier used for evaluating *soundness*, or topicality, was a multinomial Naive Bayes classifier. This is a standard classifier used in machine learning, and is typically used as a baseline for evaluation against other classifiers. The utility of the obfuscation mechanism was measured by comparing the classification accuracy of the Naive Bayes classifier on the obfuscated dataset against its performance on the unobfuscated dataset. We are interested in measuring the change in performance after obfuscation, rather than the absolute accuracy of the classifier, as an indicator of the utility of the obfuscation mechanism.

[7] Recall that the cosine similarity between two vectors is their normalised dot product.

[8] The metric properties of the Ruzicka metric can be derived from the proofs presented in [37].

For both of these metrics we will use accuracy to measure success; this is a standard measure used in machine learning and NLP.

5.4 Results

We first present results for the *safeness* and *soundness* criteria, using the authorship attribution and topic classification methods described above. Then we present a third experiment which compares the Ruzicka and WMD metrics for a subset of data points.

5.4.1 Topic Classification

We first measured how well the obfuscation mechanism preserves the topicality of the original document; recall that this is the *soundness* metric described in Section 5.1.

For the Reuters and redacted Fan fiction datasets, we established baseline topic classification accuracies (on the Raw datasets) of **81.4%** and **82.4%** respectively. We then ran the classifier over the Content-Words and BOW-n datasets using unobfuscated test data to establish a baseline, followed by obfuscation using scales 0.1, 0.2 and 0.5. The results for these runs are displayed in Table 5.3. Note that we expect a random classification accuracy of **20%** given that there are 5 topics to classify against with approximately even coverage in the training and test sets.

Dataset	**Accuracy**	**Obfuscation Accuracy**		
Reuters	Baseline	Scale=0.1	Scale=0.2	Scale=0.5
Raw	81.4	-	-	-
Content-Words	81.4	81.6	81.0	71.9
BOW-1000	80.4	80.8	80.8	75.2
BOW-500	79.2	79.4	79.4	70.7
BOW-200	76.0	76.0	76.0	66.7
BOW-50	66.3	67.9	68.1	61.7
Fan fiction	Baseline	Scale=0.1	Scale=0.2	Scale=0.5
Raw	82.4	-	-	-
Content-Words	83.3	79.4	79.4	54.9
BOW-1000	83.3	77.5	76.5	57.8
BOW-500	81.4	80.4	81.4	63.7
BOW-200	79.4	71.6	71.6	53.9
BOW-50	60.8	49.0	49.0	46.1

Table 5.3: Results for topic classification over the various unobfuscated and obfuscated test sets. Classification accuracy is significantly lower for scale=0.5, which corresponds to more obfuscation. However, accuracy is still well above the 'random' baseline of 20%.

We notice that the classification accuracy decreases as the lengths of the documents decrease. In fact, many documents in the BOW-50 test set are *empty*, meaning that they contained no words from the 50 word vocabulary chosen by the feature selector. This clearly impacts classification accuracy, and is reflected in the lower baseline accuracies for the BOW-50 datasets.

Also of note is that the obfuscation for scales 0.1 and 0.2 appeared to have a small effect on accuracy, as compared with significantly reduced accuracy at scale = 0.5. In particular, there was no change to the accuracy in the Reuters dataset, and some increase in accuracy. At scale=0.5, the results are still much better than random, showing that some topicality is clearly preserved.

5.4.2 Author Identification

We next measured how well the obfuscation mechanism protected the authorship of the output documents; this corresponds to the *safeness* criteria described earlier.

We first established baselines on the Raw datasets for Reuters and redacted Fan fiction of **71.1%** and **70.6%** respectively. As for topic classification above, we then ran the authorship identifier on non-obfuscated and obfuscated test sets, using scales of 0.1, 0.2 and 0.5. The results for these runs are shown in Table 5.4. Note that random accuracy would be **1%** for the Fan fiction dataset and **5%** for the Reuters dataset.

Dataset	**Accuracy**	**Obfuscation Accuracy**		
Reuters	Baseline	Scale=0.1	Scale=0.2	Scale=0.5
Raw	71.1	-	-	-
Content-Words	68.5	67.9	67.9	41.7
BOW-1000	65.9	62.1	63.5	41.9
BOW-500	64.1	61.7	62.1	40.9
BOW-200	47.9	46.9	48.5	27.1
BOW-50	23.9	20.0	19.0	6.2
Fan fiction	Baseline	Scale=0.1	Scale=0.2	Scale=0.5
Raw	70.6	-	-	-
Content-Words	67.7	67.7	67.6	4.9
BOW-1000	48.0	35.3	40.2	2.0
BOW-500	46.1	34.3	34.3	5.9
BOW-200	36.3	19.6	18.6	8.8
BOW-50	13.7	4.9	4.9	1.0

Table 5.4: Results for authorship attribution over the various unobfuscated and obfuscated test sets. Uniformly randomly assigning authorship would have an accuracy of 1% over 100 possible authors for the Fan fiction dataset, and 5% over 20 authors for the Reuters dataset.

These results show that the obfuscation mechanism is providing some authorship protection at scales 0.1 and 0.2, and fairly substantial protection at scale 0.5 for the Fan fiction dataset, although this needs to take into account that random guess is only 1%. For the Reuters dataset the protection is much lower, although it is not clear why this should be the case. For the Fan fiction results, the big drop in accuracy from the original dataset suggests that increasing the amount of noise further might provide closer to ‘random’ protection.

5.4.3 Analysis

The results from the topic classification and authorship attribution tests show that for scale 0.5 we see significantly higher authorship protection with some reduction in utility, particularly for the Fan fiction dataset. This suggests that the mechanism does offer some privacy against an adversary using this type of authorship identifier. Interestingly the results for the Reuters dataset suggested that the obfuscation had less impact (both on topicality and authorship protection) than for the Fan fiction dataset. This might be related to the removal of the proper nouns from the Fan fiction dataset prior to obfuscation, which could be investigated in future experiments.

It is interesting to compare the obfuscated and unobfuscated texts to see how they were modified by the mechanism, as the results seem to indicate minimal changes for scales 0.1 and 0.2. We found that the sample texts remained virtually unmodified under obfuscation with scales 0.1 and 0.2. Analysis of the noisy vectors generated indicates that the problem might be due to the sparsity of the Word2Vec space, so that noisy vectors generated using small additions of noise are still closest to the original vector. For the results with scale set to 0.5, we found that the noisy words bore little resemblance to the original words. Some examples are shown in Table 5.5. Given this disparity, it is surprising that the topic classification accuracy was so high for scale 0.5.

We suspect the noisiness of the output document is due to the use of cosine similarity by Word2Vec for distance comparison, which differs from our obfuscation mechanism (based on Laplace noise with Euclidean distance assumptions). It is possible that our alternative mechanism presented in Algorithm 1 might have more success at generating close words, as this adds noise to the angular distance using a Laplace distribution, meaning that words closer in semantic meaning may be more likely to be generated. We leave further study of this mechanism to future work.

Original	began answered prince servants king
Obfuscated	wildly diverging Caisse populaire Widianto Hendro Cahyono
Original	dwarf beard tears puppy pretty heavily murmured
Obfuscated	evangelical Christians wield refit Hi Derryn linerboard mill toasted sandwich
Original	prank bed magical realised stared weapon
Obfuscated	Sengoku twins laser refractive fulcrum snowfall sports house

Table 5.5: Some snippets of original (bag of words) text with their slightly baffling obfuscations. These obfuscations were produced using scale=0.5 on the Fan fiction dataset with 1000 features.

It would be worth investigating whether the use of Euclidean distance metrics in Word2Vec could be used to fetch semantically similar words. Whilst cosine similarity is overwhelmingly favoured in the literature as a measure of semantic distance, there have been some uses of Euclidean distance for word and document semantics; of particular note is the use of Euclidean distance in the Word

Mover's Distance metric to measure how far a word has to move from one document to another. We considered using a Euclidean distance metric to measure similarity in Word2Vec, but the computational complexity of calculating the pairwise Euclidean distances for the Word2Vec vectors made it intractable.

One of the aims of these experiments was to evaluate the feasibility of using Word2Vec to generate noisy documents whilst maintaining some semantic similarity with the original document. During experimentation we noticed that Word2Vec contains a lot of 'noisy' words, such as misspellings or unusual phrases, which occur with high cosine similarity to valid words. This affected the sensibility of documents generated using Word2Vec, and we expect this would have a bigger impact on the results if the evaluation metric included *sensibleness*, as per the original PAN author obfuscation task.

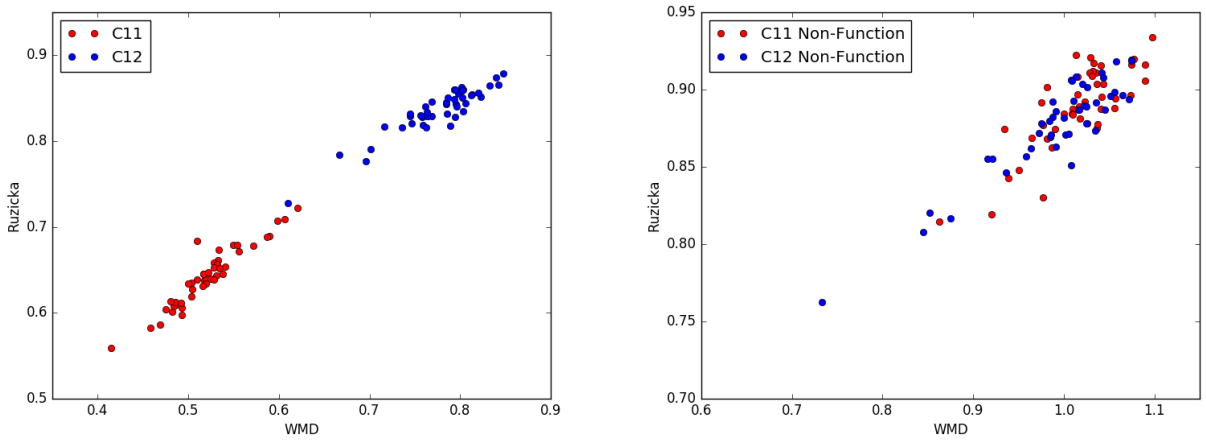
5.4.4 Comparison of Distance Metrics

We now investigate the relationship between the Word Mover's Distance and the Ruzicka metric. Recall that the Ruzicka metric is the metric of choice for the character n-gram space which best represents authors. We propose that if a simple relationship between these two metrics can be determined, then it should be possible to guarantee some degree of author-privacy given a mechanism that satisfies document-indistinguishability.

To test this hypothesis, we randomly selected sets of documents and performed pairwise distance calculations on them, using both the Word Mover's Distance and the Ruzicka metric. We used the Word Mover's Distance function provided by the gensim Word2Vec API. [9] The Ruzicka metric we used was the metric implemented in the Koppel Method codebase which we used for author attribution. We chose documents from the Reuters dataset, using the topics C11 and C12, and a selection of 10 documents from the Raw and Content-Words datasets. The results are shown in Figure 5.1.

The results suggest a linear relationship between the distance metrics. The relationship is more apparent for documents from the original dataset, although the documents from the Content-Words dataset still show linearity with higher variance. These results suggest that there could be a way to relate the document-indistinguishability and author-indistinguishability definitions from the previous chapter. Such a relationship could allow the development of a mechanism that satisfies document-indistinguishability with privacy guarantees for author-indistinguishability. This relationship is worth further investigation, which we leave to future work.

[9] https://radimrehurek.com/gensim/models/word2vec.html



(a) Raw Reuters RCV1 dataset

(b) Reuters RCV1 Content-Words dataset

Figure 5.1: Comparison of the Ruzicka and Word Mover's Distance metrics on selected Reuters documents from topics C11 and C12 showing what looks like a linear relationship.

5.5 Summary

We began this chapter with the aim of finding a relationship between the Ruzicka metric and the Word Mover's Distance. Our experiments showed that there appears to be a linear relationship between these metrics which warrants further investigation. This might enable us to infer some relationship between document-indistinguishability and author-indistinguishability in the future.

We also aimed to evaluate the use of Word2Vec to generate noisy words for document obfuscation. We experimented with two obfuscation mechanisms but found that only one could be used with the Word2Vec API chosen due to computational inefficiencies in searching. The mechanism we chose still had poor computational performance, but we were able to generate results across all of the datasets. Although the mechanism produced good results in terms of topic classification accuracy, manual inspection of the generated documents revealed many obfuscations were overly 'noisy' and appeared to be semantically unrelated to the original document. We suspected that this was a result of the mechanism design which was based on Euclidean distance assumptions. We also noted the large number of misspellings and unusual phrases in Word2Vec which may have contributed to the generation of 'noisy' documents. These experiments point to improvements which can be made to Word2Vec in order to achieve more semantically sensible documents.

Finally, we note that the mechanism which we implemented produced good results for scale 0.5, with reasonable topic classification accuracy and good author privacy as measured by the accuracy of the author identifier. It is worth investigating the use of Euclidean metrics for use with Word2Vec, as this may open the way to the use of existing Laplacian mechanisms for author obfuscation similar to the mechanisms suggested in these experiments.

6 CONCLUDING REMARKS

The motivating vision for this thesis was to develop a privacy framework for obfuscating documents so as to protect the anonymity of their authors whilst preserving the semantics of the original document. The novelty of our approach was in the use of the generalised differential privacy framework for this problem. This represents the first attempt to apply differential privacy in the text document privacy space. We simplified the problem by considering preservation of the topicality of the original document, thereby reducing the complexity of the language component of this task. We then considered both a theoretical framework and an experimental evaluation to investigate the problem, with the goals of identifying gaps in both the privacy literature and the NLP literature.

Our exploration of generalised differential privacy and geo-location privacy led us explore different metrics for use in text document privacy. We identified gaps in the literature on Laplacian mechanisms which prevented the realisation of a differentially private mechanism. In particular, the use of non-Euclidean metrics, used to measure document and word similarities in the NLP literature, is unexplored in terms of Laplacian mechanisms, which to date have only considered the use of Euclidean metrics.

Our experimentation with Word2Vec identified computational limitations with its implementation. In addition, we found that Word2Vec contained many 'noisy' words which reduced the sensibleness of the output documents. We identified these as areas for future work.

Finally, we experimentally explored the relationship between the Ruzicka metric, used to measure author distances, and the Word Mover's Distance, used to measure document distances. We found a correlation between these metrics which appears to be linear, warranting further investigation. We envisage that this relationship could be used to experimentally verify the hypothesis that document-indistinguishability implies author-indistinguishability, paving the way for mechanisms based on document-indistinguishability which can provide privacy guarantees for author obfuscation.

6.1 Discussion and Future Work

The development of this thesis presented many challenges, most particularly in finding a suitable approach to privacy in the text document domain. Whilst the intuition behind privacy for statistical databases has been well-established, thanks to the development of differential privacy, this intuition does not apply in the text document privacy domain. This may explain the plethora of approaches to privacy in this domain, and the lack of a clear direction for privacy. A number of approaches were considered before generalised differential privacy was chosen. Generalised differential privacy

presented the only privacy definition applicable to text documents which also had clear privacy guarantees. We considered this mandatory, given the demonstrated risks of relying on *ad-hoc* notions of privacy. However, the concepts underlying generalised differential privacy were challenging, and generalised differential privacy has had few applications in the literature from which to draw inspiration. The geometric intuition behind geo-indistinguishability provided some guidance for the application to author obfuscation, but the differences between these domains were not fully resolved and leave some questions still to be answered. This has opened up a number of avenues for future research.

The use of non-Euclidean metrics for our privacy definitions was a stumbling block when applying Laplacian mechanisms to our domain. The development of noise-adding mechanisms which support the use of non-Euclidean metrics is important for applications in the text document privacy domain, which is dominated by the use of non-Euclidean distance measures such as cosine similarity. In addition, the use of high-dimensional vectors is common in the text document domain, but the addition of noise to such vectors appears to be missing from the privacy literature, especially in conjunction with the use of non-Euclidean metrics. We consider the development of methods for adding Laplacian noise with these constraints to be important in opening up differential privacy to the text document domain.

Conversely, the stylometry and natural language domains have considered Euclidean distance as a similarity measure for text documents. It would be useful to investigate the use of Euclidean metrics with Word2Vec, and especially whether word embeddings could be learned in a way that is more responsive to the use of Euclidean metrics for similarity.

The development of a mechanism with a proven privacy guarantee is a future goal. We would like to consider mechanisms which will work with the Word Mover's Distance, which seems to be a natural metric for document similarity under Word2Vec. The exponential mechanism is an interesting candidate which we think would suit further investigation for the author obfuscation problem.

Finally, and more broadly, the use of differential privacy within the text document domain has been all but dismissed, but this research opens up the possibility that differential privacy could be applied to other problems in the text document privacy domain. It would be interesting to see if other problems in this domain such as text sanitisation can be reformulated to fit into the framework of generalised differential privacy.

A APPENDIX

Stopword list

The following standard stopword list is used by the Python Scikit-Learn library and was used in creating the Content-Words dataset (Section 5.3).

a, about, above, across, after, afterwards, again, against, all, almost, alone, along, already, also, although, always, am, among, amongst, amoungst, amount, an, and, another, any, anyhow, anyone, anything, anyway, anywhere, are, around, as, at, back, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, below, beside, besides, between, beyond, bill, both, bottom, but, by, call, can, cannot, cant, co, con, could, couldnt, cry, de, describe, detail, do, done, down, due, during, each, eg, eight, either, eleven, else, elsewhere, empty, enough, etc, even, ever, every, everyone, everything, everywhere, except, few, fifteen, fifty, fill, find, fire, first, five, for, former, formerly, forty, found, four, from, front, full, further, get, give, go, had, has, hasnt, have, he, hence, her, here, hereafter, hereby, herein, hereupon, hers, herself, him, himself, his, how, however, hundred, i, ie, if, in, inc, indeed, interest, into, is, it, its, itself, keep, last, latter, latterly, least, less, ltd, made, many, may, me, meanwhile, might, mill, mine, more, moreover, most, mostly, move, much, must, my, myself, name, namely, neither, never, nevertheless, next, nine, no, nobody, none, noone, nor, not, nothing, now, nowhere, of, off, often, on, once, one, only, onto, or, other, others, otherwise, our, ours, ourselves, out, over, own, part, per, perhaps, please, put, rather, re, same, see, seem, seemed, seeming, seems, serious, several, she, should, show, side, since, sincere, six, sixty, so, some, somehow, someone, something, sometime, sometimes, somewhere, still, such, system, take, ten, than, that, the, their, them, themselves, then, thence, there, thereafter, thereby, therefore, therein, thereupon, these, they, thick, thin, third, this, those, though, three, through, throughout, thru, thus, to, together, too, top, toward, towards, twelve, twenty, two, un, under, until, up, upon, us, very, via, was, we, well, were, what, whatever, when, whence, whenever, where, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, whoever, whole, whom, whose, why, will, with, within, without, would, yet, you, your, yours, yourself, yourselves

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the ACM Conference on Computer and Communications Security*, pages 308–318. (2016).

[2] Abbasi, A. and Chen, H. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7 (2008).

[3] Alvim, M. S., Chatzikokolakis, K., Degano, P., and Palamidessi, C. Differential privacy versus quantitative information flow. *CoRR*, abs/1012.4250 (2010).

[4] Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., and Palamidessi, C. Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, pages 39–54 Springer. (2011).

[5] Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P., and Si, L. t-plausibility: Generalizing words to desensitize text. *Transactions on Data Privacy*, 5(3):505–534 (2012).

[6] 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).

[7] Asghar, H. J., Tyler, P., and Kâafar, M. A. Differentially private release of public transport data: The opal use case. *CoRR*, abs/1705.05957 (2017).

[8] Bagnall, D. Author identification using multi-headed recurrent neural networks. *CoRR*, abs/1506.04891 (2015).

[9] Barbaro, M. and Zeller Jr, T. *A Face is Exposed for AOL Searcher No. 4417749*, (2006 - accessed May 24, 2017). http://www.nytimes.com/2006/08/09/technology/09aol.html.

[10] Blum, A., Ligett, K., and Roth, A. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12 (2013).

[11] Bojarski, M., Choromanska, A., Choromanski, K., and LeCun, Y. Differentially-and non-differentially-private random decision trees. *arXiv preprint arXiv:1410.6973* (2014).

[12] 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).

[13] Chaudhuri, K. and Monteleoni, C. Privacy-preserving logistic regression. In *Advances in Neural*

Information Processing Systems, pages 289–296. (2009).

[14] Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109 (2011).

[15] Chen, T., Boreli, R., Kaafar, M.-A., and Friedman, A. On the effectiveness of obfuscation techniques in online social networks. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 42–62 Springer. (2014).

[16] Cumby, C. and Ghani, R. A machine learning based system for semi-automatically redacting documents. In *Proceedings of the Twenty-Third Conference on Innovative Applications of Artificial Intelligence (IAAI)*. (2011).

[17] Dalenius, T. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15: 222–429 (1977).

[18] Dalenius, T. Finding a needle in a haystack or identifying anonymous census records. *Journal of official statistics*, 2(3):329 (1986).

[19] De Montjoye, Y.-A., Radaelli, L., Singh, V. K., et al. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539 (2015).

[20] du Pin Calmon, F. and Fawaz, N. Privacy against statistical inference. In *50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1401–1408 IEEE. (2012).

[21] Dwork, C. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)(2)*, pages 1–12. (2006).

[22] Dwork, C. and Naor, M. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):8 (2008).

[23] Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876, pages 265–284 Springer. (2006).

[24] Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407 (2014).

[25] Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333 ACM. (2015).

[26] Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., and Ristenpart, T. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security Symposium*, pages 17–32 USENIX Association. (2014).

[27] Ganta, S. R., Kasiviswanathan, S. P., and Smith, A. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 265–273 ACM. (2008).

[28] Ghosh, A., Roughgarden, T., and Sundararajan, M. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693 (2012).

[29] Hardt, M., Ligett, K., and McSherry, F. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems*, pages 2339–2347. (2012).

[30] Juola, P. *Rowling and "Galbraith": An Authorial Analysis*, (Accessed Sept 2, 2017). http://languagelog.ldc.upenn.edu/nll/?p=5315.

[31] Kacmarcik, G. and Gamon, M. Obfuscating document stylometry to preserve author anonymity. In *ACL*, pages 444–451. (2006).

[32] Karzmi, A. A. *How many authors does the Prime Minister have for his speeches: A Stylometric Analysis*, (Accessed Sept 2, 2017). http://aliarsalankazmi.github.io/blog_DA/posts/r/2016/11/18/authorial_analysis_pm.html.

[33] Kestemont, M., Stover, J., Koppel, M., Karsdorp, K., and Daelemans, W. Authorship verification with the Ruzicka metric. In *Proceedings of Digital Humanities 2016*, pages 246–249. (2016).

[34] Keswani, Y., Trivedi, H., Mehta, P., and Majumder, P. Author masking through translation. In *CLEF (Working Notes)*, pages 890–894. (2016).

[35] Koppel, M. and Winter, Y. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology JASIST*, 65(1):178–187 (2014).

[36] Koppel, M., Schler, J., and Argamon, S. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94 (2011).

[37] Kosub, S. A note on the triangle inequality for the jaccard distance. *CoRR*, abs/1612.02696 (2016).

[38] Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 957–966. (2015).

[39] Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397 (2004).

[40] Li, N., Li, T., and Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the IEEE 23rd International Conference on Data Engineering (ICDE)*, pages 106–115 IEEE. (2007).

[41] Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G. S., Hipp, J. D., Peng, L., and Stumpe, M. C. Detecting cancer metastases on gigapixel pathology images. *CoRR*, abs/1703.02442 (2017).

[42] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3 (2007).

[43] Mansoorizadeh, M., Rahgooy, T., Aminiyan, M., and Eskandari, M. Author obfuscation using wordnet and language models. notebook for PAN at CLEF 2016. In *CLEF 2016 Evaluation Labs and Workshop–Working Notes Papers. CEUR-WS. org*. (2016).

[44] McDonald, A. W., Afroz, S., Caliskan, A., Stolerman, A., and Greenstadt, R. Use fewer instances of the letter "i": Toward writing style anonymization. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 299–318 Springer. (2012).

[45] McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 94–103 IEEE. (2007).

[46] Mihaylova, T., Karadjov, G., Kiprov, Y., Georgiev, G., Koychev, I., and Nakov, P. SU@ PAN'2016: Author obfuscation. In *CLEF (Working Notes)*, pages 956–969. (2016).

[47] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781 (2013).

[48] Mosteller, F. and Wallace, D. L. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309 (1963).

[49] Muse, A. *How the NSA identified Satoshi Nakamoto*, (2017 - accessed Aug 31, 2017). https://medium.com/@amuse/how-the-nsa-caught-satoshi-nakamoto-868affcef595.

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

[51] Paganini, P. *Stylometric analysis to track anonymous users in the underground*, (2013 - Accessed Sept 2, 2017). http://securityaffairs.co/wordpress/11652/cyber-crime/stylometric-analysis-to-track-anonymous-users-in-the-underground.html.

[52] Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I. J., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. *CoRR*, abs/1610.05755 (2016).

[53] Potthast, M., Hagen, M., and Stein, B. Author obfuscation: Attacking the state of the art in authorship verification. In *CLEF (Working Notes)*, pages 716–749. (2016).

[54] Robinson, D. *Text analysis of Trump's tweets confirms he only writes the (angrier) Android half*,

(2016 - Accessed Sept 2, 2017). http://varianceexplained.org/r/trump-tweets/.

[55] Rubner, Y., Tomasi, C., and Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121 (2000).

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

[57] Sánchez, D. and Batet, M. Toward sensitive document release with privacy guarantees. *Engineering Applications of Artificial Intelligence*, 59:23–34 (2017).

[58] Sapkota, U., Bethard, S., Montes-y Gómez, M., and Solorio, T. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of NAACL: Human Language Technologies*, pages 93–102. (2015).

[59] Seidman, S. Authorship verification using the impostors method. In *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*. (2013).

[60] Stamatatos, E. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556 (2009).

[61] Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., and Stein, B. Overview of the author identification task at PAN 2014. In *CLEF (Working Notes)*, pages 877–897. (2014).

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

[63] Sweeney, L. Uniqueness of simple demographics in the US population. Technical report, Carnegie Mellon University. (2000).

[64] Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570 (2002).

[65] Yang, Y. and Pedersen, J. O. A comparative study on feature selection in text categorization. In *Proc. 14th International Conference on Machine Learning (ICML)*, volume 97, pages 412–420. (1997).