PIVOT: Privacy-preserving Outsourcing of Text Data for Word Embedding Against Frequency Analysis Attack
(Poster submission)

Yanying Li and **Wendy Hui Wang**
Department of Computer Science
Stevens Institute of Technology
Hoboken, NJ
{yli158, Hui.Wang}@stevens.edu

Boxiang Dong
Department of Computer Science
Montclair State University
Montclair, NJ
dongb@montclair.edu

Abstract
In this paper, we design *PIVOT*, a new privacy-preserving method that supports outsourcing of text data for word embedding. *PIVOT* includes a 1-to-many mapping function for text documents that can defend against the frequency analysis attack with provable guarantee, while preserving the word context during transformation.

Introduction
There has been a significant growth in the volume and variety of unstructured text data. Technologies such as text analytics and Natural Language Processing (NLP) are developed to process and analyze those massively collected data. However, the high complexity of these techniques hinder common users to perform sophisticated text analysis for their own business and research use.

In this paper, we consider an outsourcing paradigm. Alice, a data owner who has a private text document D, intends to outsource D to a third-party service provider for NLP analysis without leaking sensitive information in D. Word embedding has been used commonly in many NLP tasks. It enables machine learning techniques to process raw text data. In this paper, we assume that the server performs a prediction-based word embedding method (e.g., word2vec (Mikolov et al. 2013)) to generate the vectors of D. An important property of word embedding for NLP analysis is the context of words (i.e., the co-occurrence of words).

Apparently, a simple 1-to-1 word mapping function (i.e., each distinct, individual word is transformed to a different word) can preserve the word context in D. However, the 1-to-1 transformation function is weak against the *frequency analysis attack*, which is a common attack that can break the transformation by mapping the original and transformed words based on their frequency distribution. To defend against the frequency analysis attack, in this paper, we consider *1-to-many word mapping* transformation function: multiple instances of the same word that appear at different places of D are mapped to different words in D'.

Although the 1-to-many word mapping function can defend against the frequency analysis attack, using 1-to-many word mapping for data transformation raises the challenge of how to preserve data utility. In this paper, we consider the data utility function as the accuracy of word embedding. Finding a 1-to-many mapping that preserves the semantics of the original document D for accurate word embedding is non-trivial. As an example, consider the document in which w_1 and w_2 are of frequency 4. Out of these four occurrences, w_1 and w_2 co-occur three times. Assume w_1 (w_2 resp.) is to be replaced with s_1^1 (s_2^1 reps.) and s_1^2 (s_2^2 resp.), each of frequency 2. There are several possible schemes to replace the three (w_1, w_2) pairs. Below we we show two possible schemes:

- Scheme 1: The three (w_1, w_2) pairs are mapped to two (s_1^1, s_2^1) pairs and one (s_1^2, s_2^2) pair;
- Scheme 2: The three (w_1, w_2) pairs are mapped to one (s_1^1, s_2^1) pair, one (s_1^2, s_2^1) pair, and one (s_1^1, s_2^2) pair.

Apparently, due to the fact that the context in (w_1, w_2) pairs is preserved in two (s_1^1, s_2^1) pairs, Scheme 1 preserves the context of w_1 and w_2 better than Scheme 2.

To addresses the trade-off between privacy and utility, we design *PIVOT*, an efficient word transformation method for PrIVacy-preserving Outsourcing of Text data for word embedding. To our best knowledge, this is the first work on privacy-preserving outsourcing of text data that considers the accuracy of word embedding as the utility goal.

Preliminaries
In this paper, we use the notations shown in Table 1.

Table 1: Notations

Symbol	Notation
D / D'	document before/after transformation
w_i	a word in D
S_i	replacement candidate words of w_i
s_i^p	the p-th replacement word of word w_i

Data Utility. In this paper, we consider the quality of word embedding as the utility of the outsourced data. We measure the accuracy loss of word embedding as the change of distance between the vectors that are generated from D and D'. Formally, given two vectors v and v', the distance of v and v' is measured as following, where $cos(v, v')$ measures the cosine similarity of v and v':

$$dist(v, v') = 1 - cos(v, v') = 1 - \frac{v \cdot v'}{||v|| \cdot ||v'||}.$$

Let V and V' be the set of vectors generated from D and D'. We measure the *distance* between V and V' as the absolute difference between the sum of the cosine distance of all vector pairs in V and V'.

$$dist(D, D') = \sum_{w_i, w_j \in W, s_i^p \in S_i, s_j^q \in S_j} |dist(w_i, w_j) - dist(s_i^p, s_j^q)|$$

Our utility goal is to minimize the distance between the original document D and its transformed document D'.

Attack Model. In this paper, we only consider the frequency analysis attack. We assume the attacker may obtain the prior knowledge of word frequency distribution of D. Then by collecting the word frequency information in the transformed document D', the attacker can perform the *frequency analysis attack*, a well-known attack, to map the words in the transformed document D' back to the original words in D using their frequency distribution (Naveed, Kamara, and Wright 2015).

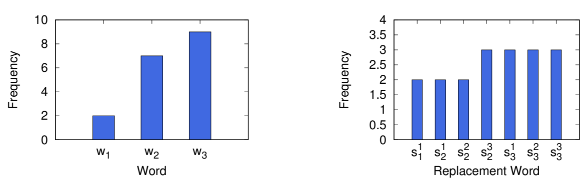
Privacy Model. To quantify the privacy guarantee against the frequency analysis attack, we define the notation of *ℓ-privacy*, which is adapted from a well-known notion ℓ-diversity (Machanavajjhala et al. 2006). Formally, given D and D', D' satisfies ℓ-privacy if for each replacement word $s \in D'$, the attacker's attack probability on s must satisfy:

$$P(s \rightarrow w) \leq \frac{1}{\ell},$$

where $w \in D$ is the original word of s , and $\ell > 1$ is a user-specified integer. Intuitively, the higher ℓ is, the stronger the transformation is against the frequency analysis attack.

Grouping-based One-to-Many Word Mapping

We design a grouping-based 1-to-many word mapping (GOMM) method that maps multiple instances of a single word to different words. To defend against the frequency analysis attack, the frequency distribution of the transformed document is always (almost) uniform. Figure 1 shows an example of before and after mapping.



(a) Freq. of orig. words (before mapping)

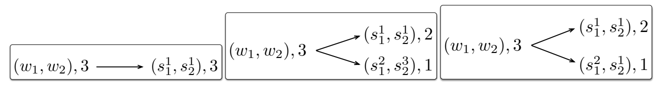
(b) Freq. of transformed words (after mapping)

Figure 1: An example of 1-to-many word mapping

To ensure that the transformation preserves the accuracy of word embedding, GOMM respects the context of the original words. For simplicity, we only consider the context that contains a single word (i.e., the immediate neighbor in the document). The key idea of GOMM is that its word replacement is based on pairs of words instead of a single word. For a given word pair (w_i, w_j), there are three possible cases of its replacement:

- Case 1 [Unique replacement]: all (w_i, w_j) pairs are replaced with the same (s_i^p, s_j^q) pair.
- Case 2 [Disjoint replacement]: the (w_i, w_j) pairs are replaced with more than one replacement pair. These replacement pairs do not overlap (i.e., a unique replacement word s_i^p is only paired with one s_j^q).
- Case 3 [Overlap replacement]: the (w_i, w_j) pairs are replaced with multiple pairs that overlap (i.e., a unique replacement word s_i^p is paired with at least two different replacement words).

Figure 2 shows an example of the three cases.



(a) Case 1 (b) Case 2 (c) Case 3

Figure 2: An example of word replacement

Apparently, for Case 1, the context of (w_i, w_j) is well preserved. Thus the utility loss of embedding of the replacement words s_i^p (s_j^q, resp.) is minimized. For Case 2, since the disjoint replacement pairs are considered as unique word pairs in the transformed document D', the context of (w_i, w_j) in D is *split* into multiple non-overlapping new contexts in D'. Thus the utility loss of Case 2 is worse than Case 1. Case 3 has the worst utility loss, as the overlapped replacement pairs make the same replacement word to be included in different contexts.

To address the trade-off between privacy and utility, we design the GOMM algorithm that consists of two steps:

- **Grouping**: all words in D are partitioned into several disjoint groups based on their frequency, with each group of size at least ℓ.
- **1-to-many mapping within groups**: for each group, pairs of words are replaced by a 1-to-many mapping function, requiring that all replacement words are of similar frequency.

The key idea of grouping is to avoid a large number of distinct replacement words, as too many distinct replacement words will create new contexts (as in Case 3) in the transformed document; the old context thus is ruined.

Conclusion

In this paper, we present our design of *PIVOT*, a privacy-preserving transformation method on raw text data that can defend against the frequency analysis attack. Next, we will implement the algorithm and perform empirical study on the real-world datasets to evaluate the performance of *PIVOT*.

References

Machanavajjhala, A.; Gehrke, J.; Kifer, D.; and Venkitasubramaniam, M. 2006. ℓ-diversity: privacy beyond κ-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, 24–24.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Naveed, M.; Kamara, S.; and Wright, C. V. 2015. Inference attacks on property-preserving encrypted databases. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 644–655. ACM.