Evaluating Differential Privacy Approaches for Query Obfuscation in Information Retrieval*

Discussion Paper

Guglielmo Faggioli[1], Nicola Ferro[1]

[1]University of Padova, Padova, Italy

Abstract

Protecting the privacy of a user while they interact with an Information Retrieval (IR) system is crucial. This becomes more challenging when the IR system is not cooperative in satisfying the user's privacy needs. Recent advancements in Natural Language Processing (NLP) have demonstrated Differential Privacy's (DP) effectiveness in safeguarding text privacy for tasks like spam detection and sentiment analysis, even under the assumption of a non-cooperative system. Our investigation explores if DP methods, originally designed for specific NLP tasks, can effectively obscure queries in IR. Our analyses show that using the Vickrey DP mechanism, employing the Mahalanobis norm with a privacy budget ranging from $\epsilon = 10$ to 12.5, provides cutting-edge privacy protection and enhances effectiveness. Unlike previous methods, DP allows users to fine-tune their desired level of privacy by adjusting the privacy budget ϵ. This flexibility offers a balance between how effective the system is and how much privacy is maintained, unlike the more rigid nature of previous approaches.

1. Introduction

Information Retrieval (IR) systems are a commodity used for many tasks, including searching for personal information, such as symptoms and diseases, political opinions, or egosurfing. Such searches can be used to profile the user and can put at risk their privacy. For example, an insurance company might try to access the user's queries to determine if they have any disease, or a malicious employee of a search engine might access the query log to blackmail them. To alleviate this, obfuscation approaches hide the sensitive information need by breaking it down into multiple non-sensitive queries. To this end, some approaches rely on replacing words with generalizations, i.e., hypernyms [2]. Other strategies use a local corpus to determine which words, by co-occurring in the documents with those in the query, induce the same ranked list [3]. We investigate for the first time whether Differential Privacy (DP) mechanisms, originally designed for specific Natural Language Processing (NLP) tasks, can effectively be used in IR to obfuscate queries. DP [4] is a state-of-the-art framework meant to release privately sensitive information. The general idea is to use a randomized mechanism that introduces noise into the computation. Thanks to this, the user can "plausibly deny" the output: it is impossible to prove that the output corresponds to the input of the user and is not due to

IRCDL 2024: 20th conference on Information and Research science Connecting to Digital and Library science, 22–23 February 2024, Bressanone, Brixen, Italy

*This is an extended abstract of [1].

the randomness of the mechanism. DP is particularly effective in the NLP domain. A line of research [5, 6, 7] operationalizes DP to release text by obfuscating each word individually. Such mechanisms work as follows: i) each word in the text is mapped to a non-contextual embedding space; ii) the embeddings are perturbed with noise drawn from a specific distribution; iii) each word is replaced with the word closest to the noisy embedding. A major advantage of DP is that it allows setting the privacy budget based on the needs of the user. This is different from current obfuscation mechanisms in IR, which are either active or not and cannot be tuned based on the user's needs. In this work, we focus on three of DP mechanisms: the Calibrated Multivariate Perturbation (CMP) [5], the Mahalanobis [6] and the Vickrey [7]. These approaches were originally devised and tested for NLP tasks that include text classification and sentiment analysis. We assume the IR system to not preserve user privacy, and to possibly be malicious. In our use case, users are the ones concerned about their privacy. They do not want to reveal their real information needs and prefer to transmit obfuscated queries to the IR system while still retrieving relevant documents. Therefore, to operationalize our mechanism, we assume each user to locally obfuscate their query and transmit the obfuscated query, or possibly multiple queries, to the IR system instead of their real query. Our goal is to determine if the DP mechanisms introduced above can successfully obfuscate users' information needs while still retrieving relevant documents.

2. Approaches

All the approaches described in this work are based on a relaxation of classical DP, called Metric-DP. To achieve traditional DP in a metric space, an obfuscation mechanism should have an equal probability of obfuscating any pair of points as the same point, irrespective of their distance. While this grants the highest level of privacy, it also requires high levels of noise, decreasing the utility of the data. In the case of metric spaces, it is often sufficient if the probability of obfuscating two points with the same one is proportional to the distance between the two points. Alternatively, the proportion of sampling a certain noise is inversely proportional to the norm of the noise itself. To this end, a relaxation of DP, called Metric-DP, has been introduced. Metric-DP [8, 9, 10] is defined as follows: given a privacy budget ϵ and a distance measure $d : \mathbb{R}^p \times \mathbb{R}^p \rightarrow [0, \infty)$, a randomized mechanism $\mathcal{M} : \mathbb{R}^p \rightarrow \mathbb{R}^p$ defined over a geometric space is Metric-DP iff, for any three points in the space $w, w', \hat{w} \in \mathbb{R}^p$, the following holds: $\frac{Pr\{\mathcal{M}(w)=\hat{w}\}}{Pr\{\mathcal{M}(w')=\hat{w}\}} \leq \exp(\epsilon d\{w, w'\})$ If the $d\{w, w'\}$ is small, w and w' are more likely to be obfuscated with the same point. Vice-versa, far apart points might be obfuscated with different points, without violating privacy constraints.

We describe here the three major DP efforts for obfuscating text in the NLP scenario, which we evaluate for the IR task. More in detail, these approaches take as input a sequence of words. Each word is mapped into a non-contextual embedding, such as GloVe [11]. Then, the embedding is obfuscated by adding some appositely sampled noise to it. To ensure that Metric-DP is achieved, the noise vector z is expected to be sampled from a distribution f such that the probability of observing z is $f(z) \propto \exp(-\epsilon||z||)$, i.e., the probability of sampling a noise with norm $||z||$ is inversely proportional to $||z||$. Finally, the closest word to the noisy embedding is used to obfuscate the corresponding word in the original text. We propose to use

these approaches in the IR scenario to perturb the queries instead of the documents, as done for NLP tasks.

The *Calibrated Multivariate Perturbation (CMP)* mechanism, defined by Feyisetan et al. [5], is based on sampling a noise vector for each term in the query following an n-dimensional Laplace distribution. Such sampling works by sampling two vectors: i) an n-dimensional unitary vector $p \in \mathbb{R}^n$ that represents the direction of the perturbation. ii) the radius of the perturbation $r \in \mathbb{R}^+$ is sampled from a Gamma distribution. To sample p, a vector $N \in \mathbb{R}^n$ is sampled from a multivariate normal distribution, with location 0 and identity covariance matrix $\mathbf{I}_n$: $N \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. Then $p = N/||N||_2$. The radius r of the noise is sampled from a Gamma distribution with shape n and scale $\frac{1}{\epsilon}$ as $r \sim Gam(n, \frac{1}{\epsilon})$. It is possible to observe that, the larger the privacy requirement, i.e., the smaller the ϵ, the bigger the noise. The noise z is defined as $z = p \cdot r$. To perturb a word w, the noise vector z is added to the original word embedding $\phi(w) \in \mathbb{R}^n$, and the word closest to the noisy word embedding is used as obfuscation. Feyisetan et al. [5] demonstrate that for any word sequence $\mathcal{W}^l$ of length $l \geq 1$ and any $\epsilon > 0$, CMP satisfies ϵd-privacy with respect to d, where d is the Euclidean distance.

The second mechanism investigated is the *Mahalanobis (Mhl)* mechanism. Xu et al. [6] noticed how the perturbation induced by CMP mechanism tends to be weak, especially for high ϵ. They hypothesize that sampling the direction of the perturbation on a circumference ($||p||_2 = 1$) increases the risk of sampling a point on an empty region. Therefore, Xu et al. adapt the CMP mechanism by transforming the direction of the noise from a circumference to an ellipsis whose orientation can be set to be towards the other embeddings. To do so, it is necessary to modify the sampling mechanism, so that, instead of sampling p such that $||p||_2 = 1$, p is sampled so that $||p||_M = 1$ where $||\cdot||_M$ is the Mahalanobis norm. To ensure that the noise z is sampled such that its probability distribution is $f(z) \propto \exp(-e||z||_M)$ a vector N is sampled from the multivariate normal distribution $N \sim \mathcal{N}(\mathbf{0}, I_n)$. Then, p is such that $p = \Sigma^{1/2} \cdot (N/||N||_2)$, where $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix of all the word embeddings. This forces the noise towards more populated areas. The sampling of the norm of the noise r is the same as for CMP.

Finally, we investigate the *Vickrey (Vkr)* mechanism. The Mhl still tends to obfuscate a word with itself for large ϵ. To reduce the probability of masking a token with itself, Xu et al. [7] define the Vickrey DP mechanism (we refer to it as Vkr). Vkr is based on two steps. In the first step, a noisy vector is sampled using any of the mechanisms described above: we can instantiate Vkr with either Mhl mechanism (Vkr_{Mhl}) or the CMP mechanism (Vkr_{CMP}). In the second step, with probability Pr the word corresponding to the closest embedding to the noisy vector is used as the obfuscation word. Vice versa, with probability $1 - Pr$ the word corresponding to the second closest embedding is used as obfuscation. The probability Pr is defined as $Pr(t, \hat{v}) = \frac{(1-t)||\phi(u_2)-\hat{v}||_2}{t||\phi(u_1)-\hat{v}||_2+(1-t)||\phi(u_2)-\hat{v}||_2}$, where $\phi(u_1)$ and $\phi(u_2)$ are respectively the closest and second closest word embeddings to $\hat{v}$, the perturbed embedding of w, and t is an additional free parameter. We set $t = 0.75$, being the best performing [7].

3. Evaluation

We consider two different collections TREC Robust '04 and TREC Deep Learning (DL '19). As word embeddings, we used GloVe [11] with 300 dimensions trained on the Common Crawl.

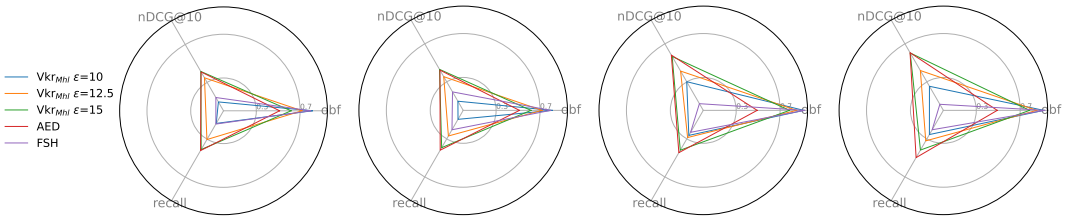
Table 1
Average MiniLM sentence similarity between the original query and 20 obfuscation queries generated with different approaches.

	Robust '04									DL '19								
ϵ	1	5	10	12.5	15	17.5	20	50	No DP	1	5	10	12.5	15	17.5	20	50	No DP
CMP	0.074	0.100	0.396	0.672	0.871	0.961	0.987	0.996		0.024	0.032	0.214	0.458	0.681	0.824	0.903	0.952	
Mhl	0.077	0.095	0.244	0.427	0.627	0.794	0.907	0.996		0.020	0.034	0.119	0.241	0.427	0.610	0.750	0.951	
Vkr$_{CMP}$	0.077	0.100	0.278	0.412	0.511	0.578	0.622	0.760		0.028	0.032	0.137	0.211	0.308	0.372	0.413	0.565	
Vkr$_{Mhl}$	0.076	0.096	0.188	0.282	0.382	0.472	0.533	0.746		0.023	0.026	0.084	0.149	0.215	0.284	0.333	0.553	
AED									0.487									0.509
FSH									0.203									0.077

In terms of retrieval models, we consider a sparse bag-of-word model, BM25, and a dense bi-encoder, Contriever [12]. To set a baseline, we compare the DP approaches aforementioned with two non-DP obfuscation approaches originally devised explicitly for the IR task. We take into consideration the seminal work by Arampatzis et al. [2], labeled AED, and the recent state-of-the-art solution by Fröbe et al. [3], labeled FSH. For each approach and for each query, we generate 20 obfuscation queries.

Table 1 shows, as a proxy of the privacy achieved by the mechanisms, the average similarity between the original query and the obfuscation queries generated to hide it. We compute the similarity as the dot product between the MiniLM [13] representations of the original query and the obfuscated ones. As expected from a DP mechanism, the higher the ϵ the higher the similarity between the queries – with $\epsilon = 50$ for both Robust '04 and DL '19, CMP and Mhl achieve similarity higher than 95%. This indicates that overall the generated queries are almost identical to the original ones and there is no substantial privacy protection. FSH, which explicitly removes synonyms and hypernyms from the queries, is particularly safe and corresponds to a DP Vkr$_{CMP}$ mechanism with $\epsilon \in [5, 10]$ or a Vkr$_{Mhl}$ with $\epsilon \in [10, 12.5]$ for the Robust '04, and DP Vkr$_{CMP}$ and a Vkr$_{Mhl}$ mechanism with $\epsilon \in [5, 10]$ for the DL '19. The privacy achieved by AED can be achieved with ϵ in the range $[10; 12.5]$ by CMP and Mhl on both collections. ϵ values that grant a comparable level of privacy are much higher for Vkr-based mechanisms, especially Vkr$_{Mhl}$, on both collections – this means that the Vkr mechanisms are substantially more secure from a privacy perspective.

As both CMP and Mhl are less effective from a privacy perspective, we focus the following analyses on the Vkr mechanism, with $\epsilon \in \{10, 12.5, 15\}$. More in detail, we compare these DP mechanisms with AED and FSH, based on three axes: i) the *obfuscation*; ii) the pooled recall; iii) the nDCG@10 observed if we re-rank the documents pooled by the obfuscation queries. We define the *obfuscation* as 1 minus the similarity of the original query and the obfuscated one. The pooled recall is obtained by transmitting to the IR system 20 obfuscated queries: for each ranked list in response to an obfuscated query, we select the first 100 documents and merge all the sets of documents obtained. We compute the recall on this new set of documents. Finally, to compute nDCG@10, we rerank the pooled documents using a different IR model (we use TAS-B to avoid biasing toward any IR model) and evaluate the quality of this ranked list. For each approach, these measures are reported on a radar plot where, as a rule of thumb, a larger area corresponds to more desirable results. Figure 1 reports the radar plots, showing the performance of different obfuscation approaches over the three axes mentioned above. We notice that the area corresponding to the AED approach (in red) is encompassed within the area corresponding to Vkr$_{Mhl}$ with $\epsilon = 15$ (green). In fact, on the Robust '04 collection, AED



(a) Robust '04, BM25 (b) Robust '04, Cnt. (c) DL '19, BM25 (d) DL '19, Cnt.

Figure 1: Performance of different obfuscation mechanisms over three axes: pooled recall, nDCG@10 of the reranked documents, obfuscation (obf), measured as 1-similarity. Cnt. stands for "Contriever".

achieves nDCG@10 of 0.410 and 0.424 for BM25 and Contriever respectively, recall of 0.420 and 0.419, and obfuscation of 0.513. Vice versa Vkr$_{Mhl}$ with $\epsilon = 15$ obtains nDCG@10 of 0.416 and 0.431, recall of 0.493 and 0.462, and obfuscation of 0.618. The exception is DL '19 with Contriever as the IR system, where AED has higher recall than Vkr$_{Mhl}$ (0.497 against 0.418). Nevertheless, this larger recall does not correspond to much larger nDCG@10, indicating that Vkr$_{Mhl}$ is preferable over AED, as it has comparable nDCG@10 (0.604 for Vkr$_{Mhl}$ against 0.607 for AED), with improved obfuscation (0.785 against 0.491). When it comes to FSH (purple), the behaviour depends on the collection. In the DL '19, using Vkr$_{Mhl}$ with $\epsilon = 10$ (blue) provides an edge over FSH: they have comparable obfuscation (0.916 the former, 0.923 the latter), but Vkr$_{Mhl}$ has much larger nDCG@10 (0.254 compared to 0.064). On the Robust '04 collection, to observe an improvement in terms of nDCG@10, it is necessary to use Vkr$_{Mhl}$ with $\epsilon = 12.5$ (nDCG@10 of 0.349 and 0.355 for BM25 and Contriever respectively) to overcome FSH in terms of nDCG@10 (0.140 and 0.194). While Vkr$_{Mhl}$ with $\epsilon = 12.5$ exhibits nDCG@10 performance slightly lower than AED, it also has obfuscation (0.719) relatively close to FSH, which has obfuscation of 0.797, much closer than AED, with obfuscation 0.513. As a general guideline, we propose to use Vkr$_{Mhl}$ as the obfuscation mechanism, with ϵ chosen in the interval [10, 15], depending on the optimal trade-off between privacy and effectiveness, as chosen by the user.

4. Conclusion and Future Work

In this work, we analyzed for the first time the performance of three DP mechanisms, originally designed for NLP, in the proxy query obfuscation IR task. We evaluated these mechanisms on the IR setting by considering three aspects: their obfuscation capabilities, their effectiveness in terms of recall, and their ability in allowing to retrieve highly relevant documents. Our findings highlight that the Vickrey mechanism with $\epsilon \in [10, 12.5]$ achieves higher privacy guarantees, with improved effectiveness, than current state-of-the-art approaches. Furthermore, lower or higher levels of ϵ allow for better satisfy the user, either in terms of privacy or accuracy, depending on their inclinations. As a future work, we plan to investigate how to perturb dense representations of the queries and combine them with generative language models to produce obfuscation queries with the same dense representation, but different terms.

References

[1] G. Faggioli, N. Ferro, Query Obfuscation for Information Retrieval through Differential Privacy, in: Procs. of ECIR 2024, 2024.

[2] A. Arampatzis, P. S. Efraimidis, G. Drosatos, Enhancing deniability against query-logs, in: Procs. of ECIR 2011, 2011, pp. 117–128.

[3] M. Fröbe, E. O. Schmidt, M. Hagen, Efficient query obfuscation with keyqueries, in: Procs. of WI-IAT '21, 2021, pp. 154–161.

[4] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. 9 (2014) 211–407.

[5] O. Feyisetan, B. Balle, T. Drake, T. Diethe, Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations, in: Procs. of WSDM '20, 2020, pp. 178–186.

[6] Z. Xu, A. Aggarwal, O. Feyisetan, N. Teissier, A differentially private text perturbation method using regularized mahalanobis metric, in: Procs. of the Second Workshop on Privacy in NLP, 2020.

[7] Z. Xu, A. Aggarwal, O. Feyisetan, N. Teissier, On a utilitarian approach to privacy preserving text generation, CoRR abs/2104.11838 (2021).

[8] M. E. Andrés, N. Bordenabe, K. Chatzikokolakis, C. Palamidessi, Geo-indistinguishability: differential privacy for location-based systems, in: A. Sadeghi, V. D. Gligor, M. Yung (Eds.), 2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013, ACM, 2013, pp. 901–914. doi:10.1145/2508859.2516735.

[9] K. Chatzikokolakis, M. Andrés, N. Bordenabe, C. Palamidessi, Broadening the scope of differential privacy using metrics, in: E. D. Cristofaro, M. K. Wright (Eds.), Privacy Enhancing Technologies - 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings, volume 7981 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 82–102. URL: https://doi.org/10.1007/978-3-642-39077-7_5. doi:10.1007/978-3-642-39077-7_5.

[10] P. Laud, A. Pankova, M. Pettai, A framework of metrics for differential privacy from local sensitivity, Proc. Priv. Enhancing Technol. 2020 (2020) 175–208. doi:10.2478/popets-2020-0023.

[11] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Procs. of EMNLP 2014, 2014, pp. 1532–1543.

[12] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, Trans. Mach. Learn. Res. (2022).

[13] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, in: Procs. of NeurIPS '20, 2020.