Measuring Actual Privacy of Obfuscated Queries in Information Retrieval

Francesco Luigi De Faveri[1 [0009−0005−8968−9485]], Guglielmo Faggioli[1 [0000−0002−5070−2049]], and Nicola Ferro[1 [0000−0001−9219−6239]]

Department of Information Engineering, University of Padova, Padova, Italy
francescoluigi.defaveri@phd.unipd.it
guglielmo.faggioli@unipd.it
nicola.ferro@unipd.it

Abstract. Privacy is a fundamental right that could be threatened by Information Retrieval (IR) models when applied and trained on sensitive data and personal user information. Although mechanisms have been proposed to protect user privacy, the effectiveness of the privacy protections is typically assessed by studying the relations between performance and parameters of the mechanisms, such as the privacy budget in Differential Privacy (DP). This often causes a disconnection between formal privacy and the privacy experienced by the user, the actual privacy. In this paper, we present the Query Inference for Privacy and Utility (QuIPU) framework, a novel evaluation paradigm to assess actual privacy based on the risk that an *"honest-but-curious"* IR system can infer the original query from the obfuscated queries received. QuIPU represents the first attempt at measuring actual privacy for IR tasks beyond the comparison of formal privacy parameters. Our analysis shows that formal privacy parameters do not imply actual privacy, causing scenarios where, for the same privacy parameter values, two systems provide different utility, but also different actual privacy. Therefore, there is a necessity for a proper way of assessing the risk, represented by QuIPU.

Keywords: Evaluation Measures · Differential Privacy · Information Retrieval · Information Security · Privacy Risks.

1 Introduction

Privacy is an essential right guaranteed by Article 12 of the Fundamental Declaration of Human Rights, which states, "No one shall be subjected to arbitrary interference with his privacy [...]". *Natural Language Processing (NLP)* models and *Information Retrieval (IR)* systems are developed using large textual datasets, including queries, documents, reviews, and online posts, frequently containing sensitive and personal user information. Including personal information in these texts, such as user profiles, personal opinions and information needs, could raise significant privacy concerns for individuals interacting with such systems. The privacy threats may compromise user safety following text analysis if

not adequately addressed. Specifically, from the examination of browser search history and retrieved documents, malicious actors can unveil sensitive details, such as an individual's salary or medical conditions [4, 15]. Heuristics strategies [3, 26] have been studied for providing privacy for IR tasks. At the same time, progress in the field of NLP has shown the potentialities of *Differential Privacy (DP)* [21] in releasing privacy-preserving text for different purposes, spanning from text classification [25], authorship anonymization [6], and query obfuscation [22]. In the IR field, DP has emerged as the leading framework for safely obfuscating user queries.

Obfuscating a query means that the real information need of a user is protected in such a way that the obfuscated queries produced can still retrieve relevant documents yet not (fully) disclose that information need. For example, the query "`how cancer grows`" may be transformed into the obfuscated alternatives "`how myeloma grows`", "`how disease spreads`", "`how melanoma evolves`". Focusing on the mechanisms' privacy parameters represents a naïve way to evaluate privacy. To this end, several attempts to assess the privacy provided have been proposed by adapting information security measures based on entropy [11, 43], and syntactic and semantic similarities [31, 55] between original and obfuscated texts. However, all these measures fall short in assessing the actual privacy achieved by the mechanism [18, 23, 35]. Indeed, in the previous example, an adversarial system could easily infer the actual user information need from the obfuscated queries "`how myeloma grows`", "`how disease spreads`", "`how melanoma evolves`", i.e. the user is looking for information about tumours, by using available query logs on its side and generate potential guesses of the original query. Yet, some value of the privacy budget would lead to generating such obfuscated queries, giving formal and proven guarantees of privacy, and most of the state-of-the-art measures would consider those queries as properly obfuscated. Moreover, since such queries are still semantically similar to the original, they would probably deliver good retrieval performance, giving a somewhat false impression of a very effective yet properly private approach. By only studying the impact of the formal mechanism parameters, there is no assurance of the risk [19] the user faces by submitting the obfuscated queries.

In this paper, we introduce the *Query Inference for Privacy and Utility (QuIPU)* framework, a novel evaluation paradigm developed to evaluate the trade-offs between the actual privacy against potential information leakage in an obfuscation protocol represented by the inference of the original query committed by a malicious IR system and the utility gained by a user. QuIPU is based on the family of attacks known as *Membership Inference Attack (MIA)* [45], adapted to a query obfuscation protocol, namely *Query Inference Attack (QuIA)*, and used against the obfuscated queries submitted to the system. It considers the potential risk that the original query is successfully inferred by a IR system after analyzing the alternative queries received and obfuscated using different configurations, i.e., using different formal privacy parameters, of an obfuscation mechanism. The measure considers the privacy-utility trade-off beyond the configuration parameters of the obfuscated mechanisms by computing a modified version of *Area*

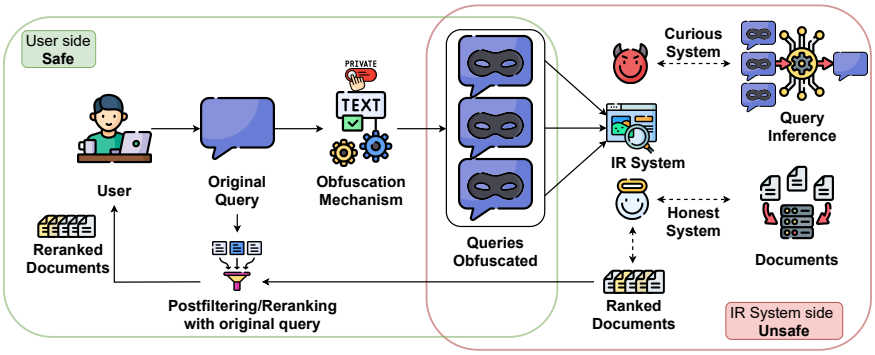


Fig. 1. Overview of the query obfuscation protocol in the presence of an *"honest-but-curious"* IR system. On the safe user side, the obfuscation mechanism takes in input the user query and produces N obfuscated queries. Such obfuscated queries are submitted to the *honest* IR System to retrieve documents. Then, if *curious*, the IR system might use the obfuscated queries to infer the original information need.

Under the Curve (AUC) of the risk *vs.* utility curve obtained. Therefore, the main contributions of this paper are: i) improving the privacy analysis paradigm by distinguishing between traditional and actual privacy guarantees within the query obfuscation process, thus studying a more comprehensive definition of privacy; ii) establishing a connection between MIAs in machine learning and its application to IR; and iii) a novel measure to assess privacy by evaluating the risk of information inference from the outputs of an obfuscation mechanism, providing insights into the formal privacy parameters.

The paper is organised as follows: Section 2 explains the context of the query obfuscation protocol; Section 3 introduces the formal definition of the QuIPU framework and the steps performed to evaluate the textual privacy provided by a mechanism. Moreover, Section 4 reports the results and discussion to investigate privacy from the *formal* privacy parameters to *actual* evaluation using the QuIPU framework. Finally, Section 5 presents the related works concerning privacy evaluation, and Section 6 draws the conclusion, outlining future directions.

2 Query Obfuscation Background

Privacy has been widely studied by the NLP and IR community [1, 32, 56, 57]. The scenario discussed in this study assumes that the users are willingly paying part of the utility during the document retrieval phase to defend the privacy of their search activity. The system is considered non-cooperative, as it does not actively contribute to protecting user privacy, e.g., it does not provide any private API to mask the information need of the user. Figure 1 illustrates the general query obfuscation protocol, focus of this work and commonly used in IR [17, 22]. The process considers two distinct environments: on the user (*safe*)

Table 1. ε-DP obfuscation mechanisms organised considering the obfuscation strategy.

Obfuscation Strategy	Mechansim	Description
Sampling	CusText [10]	Sampling of a new term is bounded to K possible terms picked using the scores computed using the distances among word embeddings.
	SanText [54]	Sampling of a new term is computed with a score based on the distances among embeddings, with terms closer to the obfuscation having a higher probability.
	TEM [7]	Noise sampled from an n - dimensional Gumbel distribution is added to the scores, and the final obfuscation term is sampled accordingly to the maximum noisy score.
	WBB [17]	Based on the word type, e.g., Nouns or Verbs, the mechanism finds the top-$(k+n)$ similar terms, excludes the first k, and samples the private word in the residuals n.
Embedding Perturbation	CMP [24]	The noise is sampled from an n - dimensional Laplace distribution of scale $\frac{1}{\varepsilon}$.
	Mhl [52]	The noise is sampled from an n - dimensional Normal distribution defined by the λ regularized Mahalanobis norm of the term embedding.
	Vickrey [53] (CMP/Mhl)	The noise is sampled as defined by the parent mechanism (CMP or Mhl) and the obfuscation term is set based on a free parameter t.

side, the original query is generated by the user and privatized using an obfuscation mechanism, i.e., an algorithm that, given an original sensitive query q, generates N non-sensitive obfuscated queries that (theoretically) prevent the unveiling of the original information need. These obfuscated queries are sent to the IR system without explicitly disclosing their information need, after the initial obfuscation process. During such operation, the user configures the mechanism parameters and privacy guarantees considering the amount of utility is lost on the downstream tasks [12, 28, 34]. On the (*unsafe*) IR system side, relevant documents are retrieved by the *"honest"* system considering the obfuscated queries received. The documents are then returned to the users for post-processing. In order to prevent a *"curious"* IR system from discovering the real user query, the obfuscation methods employed are divided into two families of mechanisms, either based on heuristics or DP strategies, summarized in the following paragraphs and in Table 1.

Heuristics Obfuscation. To protect privacy in IR tasks, non-formal obfuscation methods were proposed [3, 26]. Arampatzis et al. [3] employed the WordNet [37] database to replace original terms within the query using synonyms, hypernyms, and holonyms. The obfuscation was performed based on a hierarchical degree, i.e., the level parameter, aligned with the desired obfuscation the user desires. Such an approach was further extended by Fröbe et al. [26]. More in detail, the obfuscation approach retrieves locally the top-k documents from a local corpus. Then, using a sliding window, the sequences of n terms within such documents are taken as candidate obfuscation queries, removing those queries that contain synonyms and holonyms. To decide which query to submit to the IR system, the top-k documents retrieved on the local corpus are considered as pseudo-relevant regarding the nDCG achieved.

Differential Privacy (DP) Obfuscation. Dwork et al. [21] introduced the ε-DP framework to formalize the privacy guarantees when releasing data. Given a privacy budget $\varepsilon \in \mathbb{R}^+$, and any pair of neighbouring datasets D, D', i.e., datasets

that differ for only one entry, an obfuscation mechanism $\mathcal{M}$ is DP if it holds the inequality $\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S] \ \forall S \subset \mathrm{Im}(\mathcal{M})$. DP introduces calibrated noise levels during output computation using the privacy budget ε, which controls the balance between data privacy and utility. The adoption of the DP framework for metric spaces, and therefore for NLP tasks, has been proposed in [8]. Metric-DP extends the traditional DP definition by ensuring that the probability of obfuscating two distinct points x, x' is proportional to the distance $d(x, x')$ between them. The DP formal framework has enabled the privacy research community to propose different strategies based on noisy sampling [7, 10, 17, 54] and perturbed word embeddings [24, 52, 53], cf. Table 1.

3 The QuIPU framework

In this Section, we define the *Query Inference for Privacy and Utility (QuIPU)* framework: we report the threat model for an obfuscation protocol and the settings of QuIA. Finally, we report the risk evaluation of the attack.

3.1 Overview of the Threat Model

In this scenario, the adversary is represented by the IR system, which aims to understand the original user information need. In the query obfuscation protocol, the sweet spot for inferring the original queries is represented by the ones the system receives. The mechanism parameters, e.g., the ε privacy budget parameter of the DP obfuscation mechanisms, do not guarantee with absolute certainty that the original text is changed (or changed enough). Therefore, such queries may cause a leakage of the real information need. In addition, for the same parameters, different obfuscation strategies may produce texts with different obfuscation degrees. For instance, the effect of the parameter ε depends on the specific mechanism used [20, 33]. As a result, two DP mechanisms (one embedding-perturbation based and the other sampling-based) both parametrized with $\varepsilon = 15$ and could lead to a situation where one method achieves an actual obfuscation while the other achieves only formal obfuscation. Therefore, the IR system aims to extract as much information as possible from the received queries, previously obfuscated on the user side, and use this knowledge to infer the original text.

Consequently, the threat of a successful query inference stems not only from the obfuscation failure of the mechanism but also from the additional knowledge about the queries possessed by the adversary. The IR system might exploit its queries from the logs [9, 30, 46]: by producing a classification on the information needs carried by the obfuscated queries received and the information in its logs, it aims to improve the chances of a correct guess of the original user query. Note that if the original query is not an extremely *long tail* one [2], it is reasonable to assume that the original information need has been previously submitted to the IR system, and thus, the attack can succeed with high probability.

Finally, an important remark must be made regarding the use of cryptographic primitives in the protocol of the scenario we are analysing. Eavesdroppers or man-in-the-middle adversaries do not significantly threaten the user or the system. Cryptography can be employed while exchanging queries and documents between the client, i.e., the user, and server, i.e., the IR system, ensuring confidentiality among the internal parties of the protocol and security against external auditors. However, confidentiality does not imply privacy: if the IR system aims to disclose the user's original query, cryptography techniques alone are insufficient to safeguard privacy concerning an internal adversary.

3.2 *Query Inference Attack (QuIA)*

Algorithm 1: The *Query Inference Attack (QuIA)*.

Data: $\mathcal{Q}_{\text{obf}}$ (obf. queries q'_i), $\mathcal{Q}_{\text{logs}}$ (query log q_i), $\mathcal{T}$ (transformer encoder).
Result: Ranked list of query logs $\mathcal{R}$.
Encoding $\mathcal{T}(\mathcal{Q}_{\text{obf}}) = \{\mathcal{T}(q'_i) \in \mathbb{R}^n\}$ and $\mathcal{T}(\mathcal{Q}_{\text{logs}}) = \{\mathcal{T}(q_i) \in \mathbb{R}^n\}$;
Define $\hat{q}$ as the centroid of the vectors in $\mathcal{T}(\mathcal{Q}_{\text{obf}})$;
Compute $\mathcal{S} = [\cos(\hat{q}, \mathcal{T}(q_i)), \mathcal{T}(q_i) \in \mathcal{T}(\mathcal{Q}_{\text{logs}})]$;
Define $\mathcal{L} = [(s_i, q_i), s_i \in \mathcal{S}, q_i \in \mathcal{Q}_{\text{logs}}]$;
Sort $\mathcal{L}$ in descending order considering the similarity score s_i;
return $\mathcal{L}$;

The class of attacks known as *Membership Inference Attack (MIA)* was introduced by Shokri et al. [45] to investigate the information leakage stemming from the output of machine learning models. The attack is defined under the assumption that the attacker sees a data record but has no information about either the model parameters or the actual model architecture, i.e., a so-called *black-box* scenario. The attack is deemed successful when the attacker is able to correctly guess if the data record belongs to the model's training dataset or not.

In an obfuscation protocol, the *Query Inference Attack (QuIA)* uses the received obfuscated queries and the query logs to generate a ranked list of queries from the logs based on the similarity with the information need. Similarly to the *black-box* of the MIA scenario, the assumption is that the IR system does not know the obfuscation mechanism used on the user side and the privacy parameters of the obfuscation mechanism. Algorithm 1 reports the pseudo-code of the attack: the system receives the set of obfuscated queries $\mathcal{Q}_{\text{obf}}$ and knows its query logs $\mathcal{Q}_{\text{logs}}$. Firstly, it uses a Transformer [49] encoder $\mathcal{T}$ to obtain the embeddings of the queries in the sets[1]. Once the texts in $\mathcal{Q}_{\text{obf}}$ are encoded, it calculates the centroid $\hat{q}$ of the vectors in $\mathcal{T}(\mathcal{Q}_{\text{obf}})$, to capture the average

[1] Remark on the notation: with $\mathcal{T}(\mathcal{Q}_{\text{obf}}), \mathcal{T}(\mathcal{Q}_{\text{logs}})$ we indicate the sets of text embeddings, and with $\mathcal{T}(q'_i), \mathcal{T}(q_i)$ the singular vector embedding of the queries.

contextual similarities among the obfuscated queries received. The system computes the cosine similarity between the embeddings of the queries from the logs $\mathcal{T}(q_i) \in \mathcal{T}(\mathcal{Q}_{\text{logs}})$ and the query $\hat{q}$ to understand which queries from the logs most closely represents the average information need carried by the obfuscated queries and saves it into the list $\mathcal{S}$. The algorithm finally generates a ranked list $\mathcal{L}$ of the queries in the logs $q_i \in \mathcal{Q}_{\text{logs}}$ by sorting the pairs of (s_i, q_i) in descending order based on the similarities $s_i \in \mathcal{S}$. If the obfuscation was ineffective, then most likely, the higher a query from the logs is ranked in $\mathcal{L}$, the more likely it corresponds to the original user information need.

3.3 QuIPU Risk Modelling

Privacy is strictly linked with the definition of risk [40], i.e., the possibility that an action or event generates consequences that have an impact on what users value, in this scenario, disclosing sensitive information. The higher the risk, the lower the privacy. For example, DP obfuscation mechanisms offer the possibility that privacy and utility can be balanced by tuning the privacy budget ε. However, the framework does not provide any assurance against inference attacks [48]. To overcome this limitation, we need a formal definition of the risk against inference in the obfuscation protocol. After the QuIA algorithm has returned the ranked list $\mathcal{L}$, the IR system is tasked to guess the original query. This inference is based on the computed ranking, which considers the similarities between the obfuscated queries received (potentially leaking information) and the system's query logs (auxiliary knowledge for a correct guess of the original user query). At this point, the IR system strategy to guess the correct query is sequential: knowing that the first query is the most similar to the average information need carried by the obfuscated queries, it represents the best choice for the guess. If the first query in the logs $\mathcal{L}$ is the correct query, the attack is successful, and there is a 100% risk of correct inference. On the other hand, if the first one is not the correct guess, the adversary tries with the second query in the list, and so on, until the original query is guessed, decreasing the risk of success. Therefore, the risk r_t of successful QuIA in t guesses can be defined as the probability that the IR system correctly guesses $\bar{q}$ as the original query q, seeing the sets $\mathcal{Q}_{\text{obf}}$ and $\mathcal{Q}_{\text{logs}}$, i.e., $r_t = \mathbb{P}\left[\{\bar{q} = q\} \cap \{t \leq k\} \,|\mathcal{Q}_{\text{obf}}, \mathcal{Q}_{\text{logs}}\right]$, with k the maximum number of guessing attempts the IR system is willing to take. The upper bound for the value of k is determined by the size of the set $\mathcal{Q}_{\text{logs}}$. However, determining the precise threshold t and assessing the risk the user faces is impossible without access to the IR system's internal data and kind of attack. Therefore, we propose to model the malicious IR system with three kinds of attackers, representing relevant use cases: i) the "*lazy*" attacker, i.e., the one that looks only at the top position of the ranked list $\mathcal{L}$ and makes only one guess; ii) the "*active*" attacker, i.e., an adversary that selects the top-k queries and checks only with them if the guess is correct; and iii) the "*motivated*" attacker, i.e., the one that tries all the queries until the original one has been found. To model the probability of the risk a user faces against each of such attackers, we propose to use proxy indicators computed on where the original query appears in the ranked list $\mathcal{L}$: Precision at

1 ($P@1$) for the lazy attacker, Recall at k ($R@k$) for the active attacker, and *Reciprocal Rank (RR)* for the motivated attacker.

Drawing inspiration from the usual ROC AUC, Figure 2 illustrates the evaluation plane that links the risk r of a successful QuIA and the utility u measure considering a set of queries $\mathcal{Q}_{\text{obf}}(p_i)$ – an effectiveness measure such as nDCG in the IR case – obfuscated by a certain formal parameter p_i – the ε parameter in case of DP. In the risk-utility plane, the *Risk-Utility Boundary line*, i.e., the diagonal, describes two regions where the risk-utility trend $f(r, u)$ can be: i) above the line indicates that the utility u exceeds the associated risk r, and ii) below the line, where u is less than r. Therefore, we can define the QuIPU score, Equation 1, considering the different pairs (r, u) estimated by submitting to the IR system the set of obfuscated queries $\mathcal{Q}_{\text{obf}}(p_i)$.

$$\text{QuIPU} = 2(\mathcal{R}_+ + \mathcal{R}_-) = 2\int_{\mathcal{R}_+} f\,(r, u)\,d\nu + 2\int_{\mathcal{R}_-} f\,(r, u)\,d\nu \tag{1}$$

where $d\nu$ represents an infinitesimal variation on the *Risk-Utility Boundary line*, and the factor 2 is introduced to map the score from $\left[-\frac{1}{2}, \frac{1}{2}\right]$ to $[-1, 1]$ interval. The integrals are calculated with respect to the diagonal of the plane, such that regions where the curve lies below this diagonal, i.e., R_-, are assigned negative values, indicating that the risk r is greater than the utility u. Conversely, positive values are computed for regions where the utility u exceeds the risk r, i.e., R_+. In Figure 2, four critical points are defined:

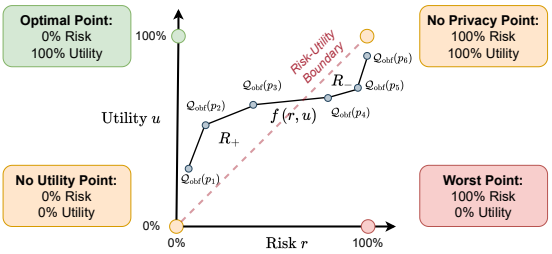


Fig. 2. Risk r *vs.* Utility u model to evaluate actual privacy. The four labels describe the plane's critical points, and the red dashed line shows the Boundary tracing two areas, R_+ and R_-, where the QuIPU score is positive or negative. The curve f shows the r *vs.* u trend when the system receives the queries $\mathcal{Q}_{\text{obf}}(p_i)$ masked with a parameter p_i.

– **No Utility Point**: This point shows when the risk and utility are both reduced to 0. It corresponds to the situation where the obfuscation mechanism fully modifies the original query, completely stopping a QuIA. However, the user completely renounces the effectiveness of the task, i.e., the submitted queries failed to retrieve any relevant documents from the IR system.

- **No Privacy Point**: This point illustrates the effect of not using the obfuscation protocol. The queries are not obfuscated, meaning the original query is fully exposed to the IR system, resulting in 100% risk, i.e., the attacker has all the information to infer correctly. Yet, utility is fully achieved, as the system can use the original query to retrieve all the relevant documents.
- **Optimal Point**: The optimal point is the best we can *theoretically* obtain. The obfuscation mechanism provides complete protection against Query Inference attacks, i.e., 0% of risk, while maintaining maximum utility. The user's information need are entirely met during the retrieval without exposing any information related to the original query.
- **Trash Point**: This is the opposite of the optimal point. The mechanism neither obfuscates the query nor these queries can retrieve any relevant documents. This case can happen if we change the *"honest-but-curious"* assumption to a *"fully-dishonest"* scenario, a.e., a phishing IR system [39, 42].

4 Experimental Evaluation

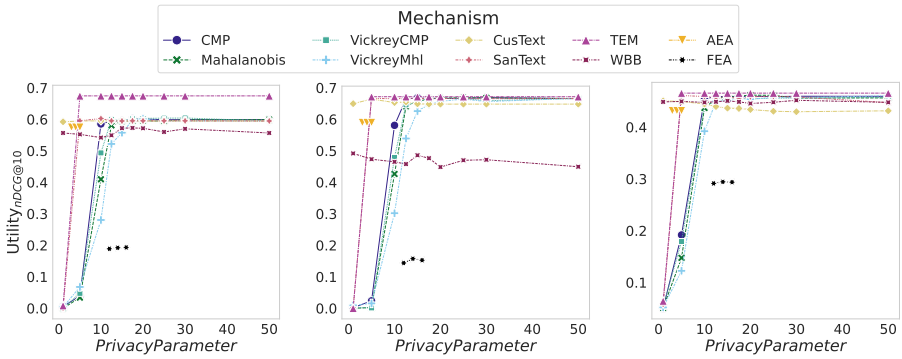
We outline here the experimental setup and compare the results observed for the traditional and actual privacy analysis using the QuIPU framework. Further analyses, the data used, and the source code are publicly available[2].

4.1 Experimental Setup

User side. We test the QuIPU framework on three different TREC collections: Deep Learning (DL'19) [14] and Deep Learning (DL'20) [13], based on the MS MARCO [41] passages corpus, containing 43 and 54 queries respectively, and the Robust '04 (Robust '04) [50] which relies on disks 4 and 5 of the TIPSTER corpus and contains 249 queries. As obfuscation mechanisms, we consider those described in Section 2, using their implementation provided by the pyPANTERA framework [16]. As privacy budget ε, we followed the parametrization reported in the original papers, which is also the one used by the pyPANTERA framework and other recent experiments [22]. In detail, we select $\varepsilon \in \{1, 5, 10, 12.5, 15, 17.5, 20, 25, 30, 50\}$. The heuristics obfuscation mechanisms, i.e., AEA [3] and FEA [26], have been parametrized using different synonyms levels $\{3, 4, 5\}$, and sliding windows sizes $\{12, 14, 16\}$, respectively. We generated 50 obfuscation variants for each query and mechanism configuration. Finally, the IR system used as re-ranker in the post-retrieval phase of the protocol is the neural dense model Contriever [29], as proposed in [17, 22].

IR system side. On the one hand, the *honest* aspect of the IR system, i.e., the part that performs the document retrieval task, is showcased by the Contriever model; on the other hand the *curious* part of the system uses as encoder model `distilbert-base-uncased` [44].We use two different models for the two tasks

[2] https://github.com/Kekkodf/QuIPU_Framework



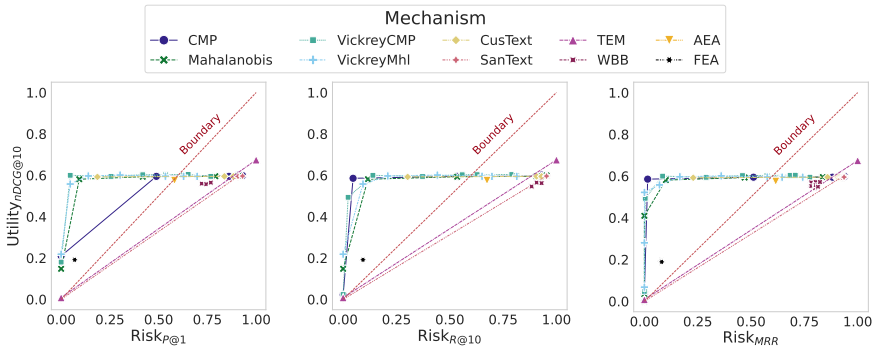
(a) DeepLearning '19 (b) DeepLearning '20 (c) Robust'04

Fig. 3. nDCG@10 (*Utility*) when varying privacy formal parameters (*PrivacyParameters*) of the obfuscation mechanisms: some mechanisms, such as TEM, achieve immediately high performance, with unclear consequences on actual privacy.

to obtain unbiased results, in line with [17, 22]. To simulate a realistic scenario for the curious IR to perform the QuIA, we use as query logs the AOL collection available in the `ir_datasets` catalog[3]. We sampled 750k queries, to which we added the original queries as explained in Section 3.

4.2 Privacy from the Formal Parameters Analysis

The traditional privacy analysis evaluates the utility as a function of the formal privacy parameters, e.g., ε. Figure 3 reports the results of the nDCG@10 *vs* the formal privacy parameters on the three different collections analysed. Note that the x-axis, representing the *PrivacyParameter*, considers both the values for the ε parameter of the DP mechanisms and also the parameters of the heuristics [3, 26]. From this traditional perspective, it emerges that with lower values of the privacy parameters, mechanisms based on a DP strategy, like TEM or SanText, achieve higher effectiveness for low values of the privacy parameter ε. On the other hand, obfuscation mechanisms based on the embedding obfuscation strategies perform with high effectiveness only if the formal parameter is high. Finally, the Heuristics show high nDCG@10 for AEA and the worst results for the FEA mechanism. These results show a misleading sense of privacy: high-performance results do not imply the actual privacy of the texts, i.e., the submitted queries are the original ones. To address this issue, we analyse the impact of obfuscation on denying a correct inference of the original query.



(a) *Lazy Attacker* (b) *Active Attacker* (c) *Motivated Attacker*

Fig. 4. Resulting Risk estimated *vs.* Utility (i.e., retrieval nDCG@10) achieved sending the $\mathcal{Q}_{\text{obf}}$ of the MSMARCO Deep Learning'19 collection. The risk is estimated as explained in Section 3, showing Risk$_{P@1}$, Risk$_{R@10}$, and Risk$_{MRR}$ for the different adversaries.

4.3 Privacy Analysis using the QuIPU Framework

In this section, we report the evaluation of the actual privacy the obfuscation mechanisms provide to the original user query. Figure 4 show the different Risk r *vs.* Utility u evaluation planes on the MSMARCO Deep Learning'19 collection obtained for the different malicious IR system attacker nature that can perform the QuIA, i.e., "*lazy*" (Figure 4(a)), "*active*" (Figure 4(b)) and "*motivated*" (Figure 4(c)). The plots show that the utility of the DP mechanisms, i.e., the nDCG@10, remains stable after an initial increase from the "No Utility Point", see Figure 2. On the other hand, the risk in all three scenarios increases, implying a high probability of discovering the real user information need from the obfuscated queries submitted. This confirms our hypothesis: although some privacy configurations give formal and theoretically proven guarantees of privacy, such obfuscated queries are vulnerable against the QuIA, even considering the "*Lazy*" attacker (Figure 4(a)), under the false impression of effective obfuscation. Conversely, the heuristics AEA and FEA mechanisms achieve the same utility and risk on all the privacy configurations. The former is always allocated around the Risk *vs.* Utility Boundary line, implying a risk, thus a probability of correct inference of the original query, equal to the utility achieved. The latter is more prone to defending user privacy against the inference of the original information need, but renouncing more than 60% of nDCG@10.

Table 2 reports the QuIPU scores obtained analysing the Risk *vs.* Utility on each set of obfuscated queries of the collections. The results show three distinct patterns that can be traced back to the three different obfuscation strategies. The Sampling-based mechanisms show weaker defences against the three attackers, and especially against the "*active*" one, i.e., QuIPU score more negative.

[3] https://ir-datasets.com/aol-ia.html

Table 2. QuIPU Score computed organizing the results by obfuscation strategy. computed using the Equation 1, measured in terms of u = nDCG@10, and the risk r of a successful Query Inference considering different adversary models. Positive values correspond to a better Utility-Privacy trade-off, cf. Section 3.

Obfuscation Strategy	Mechanism	*Lazy Attacker*			*Active Attacker*			*Motivated Attacker*		
		DL'19	**DL'20**	**Robust**	**DL'19**	**DL'20**	**Robust**	**DL'19**	**DL'20**	**Robust**
Sampling	CusText	0.041	0.109	0.034	-0.014	0.010	-0.084	0.020	0.074	0.028
	SanText	-0.247	-0.222	0.046	-0.277	-0.252	-0.237	-0.255	-0.231	0.043
	TEM	-0.264	-0.274	0.028	-0.264	-0.274	-0.329	-0.264	-0.274	0.027
	WBB	-0.005	-0.002	0.001	-0.001	-0.022	-0.011	-0.006	-0.010	0.001
Embedding Perturbation	CMP	0.299	0.372	0.175	0.257	0.323	0.001	0.283	0.353	0.170
	Mahalanobis	0.280	0.381	0.202	0.258	0.363	0.090	0.272	0.371	0.200
	VickreyCMP	0.341	0.430	0.194	0.310	0.411	0.103	0.334	0.424	0.193
	VickreyMhl	0.342	0.426	0.199	0.318	0.410	0.119	0.335	0.421	0.199
Heuristics	AEA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	FEA	0.001	0.001	0.001	0.000	0.001	0.002	0.000	0.001	0.001

Differently, the Embedding Perturbation mechanisms that are able to defend the user information need against the attackers, achieving higher QuIPU scores even when facing the "*motivated*" attacker. This suggests that, when using as DP obfuscation mechanisms, if the user wants to achieve strong actual privacy guarantees against the QuIA, it should select an obfuscation relying on changing the word embeddings of the queries. Finally, the heuristics obfuscation strategies obtain a null QuIPU score due to the stable risk and utility achieved. FEA reaches a slightly positive QuIPU score against the three attackers, implying that it is impractical for an attacker to guess the original query even if "*motivated*" to do so.

5 Related Works

Different works have been proposed to organize available privacy measures [47, 51]. Wagner and Eckhoff [51] systematically classified over eighty privacy metrics, offering a comprehensive framework for assessing privacy across different domains, e.g., communication, databases, and social networks. The survey underscores the significance of identifying the specific aspect of privacy that a metric aims to quantify, suggesting nine guiding questions for selecting the appropriate privacy measures. Specifically, the authors underlined the importance of considering the adversary's knowledge and capability when evaluating privacy. In addition, Sousa and Kern [47] described how different mechanisms developed for NLP tasks provide privacy for textual data and which can be the threats in such scenarios. Moreover, Habernal [27] stressed the importance of not relying strictly on formal analysis of DP and its application on NLP tasks, but to push research towards a concrete measurement of the privacy provided to texts.

Traditional methods for evaluating privacy primarily focus on calculating the failure rates of obfuscation mechanisms [11] or assessing the similarities between original and obfuscated texts [22, 36]. On the one hand, uncertainty measures

such as N_w and S_w [24, 52] estimate the probability that a term w remains unchanged after obfuscation and the minimum cardinality of the set of words to which the mechanism maps w, respectively. However, such measures do not capture if the mechanism changes the original term with a closely related one. On the other hand, the similarity between the original and obfuscated texts is commonly estimated using metrics like the Jaccard index or cosine similarity between sentence embeddings computed by a Transformer model, drawing inspiration by the use of BERTScores used to evaluate the quality of generated texts [55]. Meisenbacher et al. [36] proposed the α-PUC score to compute an α weighted mean between uncertainty, similarity measures, and utility preserved. However, none of the above measures offer insights into the actual privacy afforded to the texts, nor do they assess the adversarial potential to infer the original meaning of the obfuscated text. Blanco-Justicia et al. [5] criticize the reliance on formal privacy analysis solely based on the privacy budget ε parameter. They argue that DP mechanisms employing configurations where $\varepsilon > 1$ lack a comprehensive analysis of actual privacy guarantees, raising concerns about the sufficiency of privacy protection in practice[4]. In addition, Damie et al. [15] introduced a novel indicator to assess the risk of successful query recovery attacks within searchable encryption protocols. The study revealed that, even without additional background knowledge, an adversary could reconstruct the original queries with a success rate of 85%, encouraging analysis of privacy measures employed considering real adversarial scenarios of application.

To the best of our knowledge, this study represents the first attempt to bridge the gap between the evaluation based on formal parameters and actual privacy analysis, proposing a novel measure to assess the risk of inferring the original query from the set of obfuscations produced to protect user privacy against a malicious IR system.

6 Conclusion and Future Work

Assessing the privacy guarantees provided to users during IR tasks remains an open challenge. Relying solely on a formal privacy analysis considering the mechanism parameters is insufficient for concretely evaluating the privacy of obfuscation mechanisms. In this study, we introduced the QuIPU framework, a new benchmark designed to assess actual privacy provided to queries in an obfuscation protocol. We empirically evaluated the risk that an "honest-but-curious" IR system can accurately infer the original query from the obfuscated ones received using its queries from the logs. Our findings demonstrate that strong formal privacy guarantees do not necessarily imply actual privacy protection. In future work, we plan to explore additional proxy measures, e.g., the Rank-biased precision [38], to investigate their correlation with the QuIPU score. In addition, we intend to explore the capabilities of Large Language Models in determining whether or not a query has been sufficiently obfuscated.

[4] The DP configurations with $\varepsilon > 1$ deviate from the "theoretically safe" privacy setup, i.e., strong assurance about the formal privacy introduced, see DP definition [21].

References

1. Ahmad, W.U., Chang, K., Wang, H.: Intent-aware query obfuscation for privacy protection in personalized web search. In: Collins-Thompson, K., Mei, Q., Davison, B.D., Liu, Y., Yilmaz, E. (eds.) The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, pp. 285–294, ACM (2018), https://doi.org/10.1145/3209978.3209983, URL https://doi.org/10.1145/3209978.3209983
2. Anderson, C.: The long tail. Effective Business Model on the Internet—Moscow: Mann, Ivanov & Ferber (2012)
3. Arampatzis, A., Drosatos, G., Efraimidis, P.: A versatile tool for privacy-enhanced web search. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S.M., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings, Lecture Notes in Computer Science, vol. 7814, pp. 368–379, Springer (2013), https://doi.org/10.1007/978-3-642-36973-5_31, URL https://doi.org/10.1007/978-3-642-36973-5_31
4. Bavadekar, S., Dai, A.M., Davis, J., Desfontaines, D., Eckstein, I., Everett, K., Fabrikant, A., Flores, G., Gabrilovich, E., Gadepalli, K., Glass, S., Huang, R., Kamath, C., Kraft, D., Kumok, A., Marfatia, H., Mayer, Y., Miller, B., Pearce, A., Perera, I.M., Ramachandran, V., Raman, K., Roessler, T., Shafran, I., Shekel, T., Stanton, C., Stimes, J., Sun, M., Wellenius, G., Zoghi, M.: Google COVID-19 search trends symptoms dataset: Anonymization process description (version 1.0). CoRR **abs/2009.01265** (2020), URL https://arxiv.org/abs/2009.01265
5. Blanco-Justicia, A., Sánchez, D., Domingo-Ferrer, J., Muralidhar, K.: A critical review on the use (and misuse) of differential privacy in machine learning. ACM Comput. Surv. **55**(8), 160:1–160:16 (2023), https://doi.org/10.1145/3547139, URL https://doi.org/10.1145/3547139
6. Bo, H., Ding, S.H.H., Fung, B.C.M., Iqbal, F.: ER-AE: differentially private text generation for authorship anonymization. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pp. 3997–4007, Association for Computational Linguistics (2021), https://doi.org/10.18653/V1/2021.NAACL-MAIN.314, URL https://doi.org/10.18653/v1/2021.naacl-main.314
7. Carvalho, R.S., Vasiloudis, T., Feyisetan, O., Wang, K.: TEM: high utility metric differential privacy on text. In: Shekhar, S., Zhou, Z., Chiang, Y., Stiglic, G. (eds.) Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023, pp. 883–

890, SIAM (2023), https://doi.org/10.1137/1.9781611977653.CH99, URL https://doi.org/10.1137/1.9781611977653.ch99
8. Chatzikokolakis, K., Andrés, M.E., Bordenabe, N.E., Palamidessi, C.: Broadening the scope of differential privacy using metrics. In: Cristofaro, E.D., Wright, M.K. (eds.) Privacy Enhancing Technologies - 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings, Lecture Notes in Computer Science, vol. 7981, pp. 82–102, Springer (2013), https://doi.org/10.1007/978-3-642-39077-7_5, URL https://doi.org/10.1007/978-3-642-39077-7_5
9. Chau, M., Fang, X., Sheng, O.R.L.: Analysis of the query logs of a web site search engine. J. Assoc. Inf. Sci. Technol. **56**(13), 1363–1376 (2005), https://doi.org/10.1002/ASI.20210, URL https://doi.org/10.1002/asi.20210
10. Chen, S., Mo, F., Wang, Y., Chen, C., Nie, J.Y., Wang, C., Cui, J.: A customized text sanitization mechanism with differential privacy. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, pp. 5747–5758, Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://doi.org/10.18653/v1/2023.findings-acl.355, URL https://aclanthology.org/2023.findings-acl.355
11. Clauß, S., Schiffner, S.: Structuring anonymity metrics. In: Juels, A., Winslett, M., Goto, A. (eds.) Proceedings of the 2006 Workshop on Digital Identity Management, Alexandria, VA, USA, November 3, 2006, pp. 55–62, ACM (2006), https://doi.org/10.1145/1179529.1179539, URL https://doi.org/10.1145/1179529.1179539
12. Clifton, C., Tassa, T.: On syntactic anonymity and differential privacy. In: Chan, C.Y., Lu, J., Nørvåg, K., Tanin, E. (eds.) Workshops Proceedings of the 29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013, pp. 88–93, IEEE Computer Society (2013), https://doi.org/10.1109/ICDEW.2013.6547433, URL https://doi.org/10.1109/ICDEW.2013.6547433
13. Craswell, N., Mitra, B., Yilmaz, E., Campos, D.: Overview of the TREC 2020 deep learning track. CoRR **abs/2102.07662** (2021), URL https://arxiv.org/abs/2102.07662
14. Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M.: Overview of the TREC 2019 deep learning track. CoRR **abs/2003.07820** (2020), URL https://arxiv.org/abs/2003.07820
15. Damie, M., Hahn, F., Peter, A.: A highly accurate query-recovery attack against searchable encryption using non-indexed documents. In: Bailey, M.D., Greenstadt, R. (eds.) 30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021, pp. 143–160, USENIX Association (2021), URL https://www.usenix.org/conference/usenixsecurity21/presentation/damie
16. De Faveri, F.L., Faggioli, G., Ferro, N.: py-PANTERA: A Python PAckage for Natural language obfuscaTion Enforcing pRivacy & Anonymization. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID,

USA., p. 6, Springer (2024), https://doi.org/10.1145/3627673.3679173, URL https://doi.org/10.1145/3627673.3679173
17. De Faveri, F.L., Faggioli, G., Ferro, N.: Words Blending Boxes. Obfuscating Queries in Information Retrieval using Differential Privacy. CoRR **abs/2405.09306** (2024), https://doi.org/10.48550/ARXIV.2405.09306, URL https://doi.org/10.48550/arXiv.2405.09306
18. Domingo-Ferrer, J., Sánchez, D., Blanco-Justicia, A.: The limits of differential privacy (and its misuse in data release and machine learning). Commun. ACM **64**(7), 33–35 (2021), https://doi.org/10.1145/3433638, URL https://doi.org/10.1145/3433638
19. Duncan, G., Keller-McNulty, S., Stokes, L.: Disclosure risk vs. data utility: The ru confidentiality map. A Los Alamos National Laboratory Technical Report **LA-UR-01-6428**, 1–30 (2001)
20. Dwork, C., Kohli, N., Mulligan, D.K.: Differential privacy in practice: Expose your epsilons! J. Priv. Confidentiality **9**(2) (2019), https://doi.org/10.29012/JPC.689, URL https://doi.org/10.29012/jpc.689
21. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) Theory of Cryptography, pp. 265–284, Springer Berlin Heidelberg, Berlin, Heidelberg (2006), ISBN 978-3-540-32732-5
22. Faggioli, G., Ferro, N.: Query obfuscation for information retrieval through differential privacy. In: Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I. (eds.) Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I, Lecture Notes in Computer Science, vol. 14608, pp. 278–294, Springer (2024), https://doi.org/10.1007/978-3-031-56027-9_17, URL https://doi.org/10.1007/978-3-031-56027-9_17
23. Faveri, F.L.D., Faggioli, G., Ferro, N.: Beyond the parameters: Measuring actual privacy in obfuscated texts. In: Roitero, K., Viviani, M., Maddalena, E., Mizzaro, S. (eds.) Proceedings of the 14th Italian Information Retrieval Workshop, Udine, Italy, September 5-6, 2024, CEUR Workshop Proceedings, vol. 3802, pp. 53–57, CEUR-WS.org (2024), URL https://ceur-ws.org/Vol-3802/paper5.pdf
24. Feyisetan, O., Balle, B., Drake, T., Diethe, T.: Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In: Caverlee, J., Hu, X.B., Lalmas, M., Wang, W. (eds.) Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 178–186, ACM (Jan 2020), https://doi.org/10.1145/3336191.3371856
25. Feyisetan, O., Kasiviswanathan, S.: Private release of text embedding vectors. In: Pruksachatkun, Y., Ramakrishna, A., Chang, K.W., Krishna, S., Dhamala, J., Guha, T., Ren, X. (eds.) Proceedings of the First Workshop on Trustworthy Natural Language Processing, pp. 15–27, Association for Computational Linguistics, Online (Jun 2021), https://doi.org/10.18653/v1/2021.trustnlp-1.3, URL https://aclanthology.org/2021.trustnlp-1.3

26. Fröbe, M., Schmidt, E.O., Hagen, M.: Efficient query obfuscation with keyqueries. In: He, J., Unland, R., Jr., E.S., Tao, X., Purohit, H., van den Heuvel, W., Yearwood, J., Cao, J. (eds.) WI-IAT '21: IEEE/WIC/ACM International Conference on Web Intelligence, Melbourne VIC Australia, December 14 - 17, 2021, pp. 154–161, ACM (2021), https://doi.org/10.1145/3486622.3493950, URL https://doi.org/10.1145/3486622.3493950
27. Habernal, I.: When differential privacy meets NLP: the devil is in the detail. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pp. 1522–1528, Association for Computational Linguistics (2021), https://doi.org/10.18653/V1/2021.EMNLP-MAIN.114, URL https://doi.org/10.18653/v1/2021.emnlp-main.114
28. Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B.C., Roth, A.: Differential privacy: An economic method for choosing epsilon. In: IEEE 27th Computer Security Foundations Symposium, CSF 2014, Vienna, Austria, 19-22 July, 2014, pp. 398–410, IEEE Computer Society (2014), https://doi.org/10.1109/CSF.2014.35, URL https://doi.org/10.1109/CSF.2014.35
29. Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E.: Unsupervised dense information retrieval with contrastive learning. Trans. Mach. Learn. Res. **2022** (2022), URL https://openreview.net/forum?id=jKN1pXi7b0
30. Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. Inf. Process. Manag. **36**(2), 207–227 (2000), https://doi.org/10.1016/S0306-4573(99)00056-4, URL https://doi.org/10.1016/S0306-4573(99)00056-4
31. Kang, Y., Liu, Y., Niu, B., Tong, X., Zhang, L., Wang, W.: Input perturbation: A new paradigm between central and local differential privacy. CoRR **abs/2002.08570** (2020), URL https://arxiv.org/abs/2002.08570
32. Klymenko, O., Meisenbacher, S., Matthes, F.: Differential privacy in natural language processing the story so far. In: Feyisetan, O., Ghanavati, S., Thaine, P., Habernal, I., Mireshghallah, F. (eds.) Proceedings of the Fourth Workshop on Privacy in Natural Language Processing, pp. 1–11, Association for Computational Linguistics, Seattle, United States (Jul 2022), https://doi.org/10.18653/v1/2022.privatenlp-1.1, URL https://aclanthology.org/2022.privatenlp-1.1
33. Kohli, N., Laskowski, P.: Epsilon voting: Mechanism design for parameter selection in differential privacy. In: 2018 IEEE Symposium on Privacy-Aware Computing, PAC 2018, Washington, DC, USA, September 26-28, 2018, pp. 19–30, IEEE (2018), https://doi.org/10.1109/PAC.2018.00009, URL https://doi.org/10.1109/PAC.2018.00009
34. Lee, J., Clifton, C.: How much is enough? choosing ϵ for differential privacy. In: Lai, X., Zhou, J., Li, H. (eds.) Information Security, 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011.

Proceedings, Lecture Notes in Computer Science, vol. 7001, pp. 325–340, Springer (2011), https://doi.org/10.1007/978-3-642-24861-0_22, URL https://doi.org/10.1007/978-3-642-24861-0_22

35. Mattern, J., Weggenmann, B., Kerschbaum, F.: The limits of word level differential privacy. In: Carpuat, M., de Marneffe, M., Ruíz, I.V.M. (eds.) Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022, pp. 867–881, Association for Computational Linguistics (2022), https://doi.org/10.18653/V1/2022.FINDINGS-NAACL.65, URL https://doi.org/10.18653/v1/2022.findings-naacl.65
36. Meisenbacher, S.J., Nandakumar, N., Klymenko, A., Matthes, F.: A comparative analysis of word-level metric differential privacy: Benchmarking the privacy-utility trade-off. In: Calzolari, N., Kan, M., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, pp. 174–185, ELRA and ICCL (2024), URL https://aclanthology.org/2024.lrec-main.16
37. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM **38**(11), 39–41 (1995), https://doi.org/10.1145/219717.219748, URL https://doi.org/10.1145/219717.219748
38. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. **27**(1), 2:1–2:27 (2008), https://doi.org/10.1145/1416950.1416952, URL https://doi.org/10.1145/1416950.1416952
39. Moore, T., Clayton, R.: Evil searching: Compromise and recompromise of internet hosts for phishing. In: Dingledine, R., Golle, P. (eds.) Financial Cryptography and Data Security, 13th International Conference, FC 2009, Accra Beach, Barbados, February 23-26, 2009. Revised Selected Papers, Lecture Notes in Computer Science, vol. 5628, pp. 256–272, Springer (2009), https://doi.org/10.1007/978-3-642-03549-4_16, URL https://doi.org/10.1007/978-3-642-03549-4_16
40. National Institute of Standards and Technology: Information security. Tech. Rep. National Institute of Standards and Technology Special Publication 800-60, Volume 1 Revision 1, August, 2008, U.S. Department of Commerce, Washington, D.C. (2008), https://doi.org/https://doi.org/10.6028/NIST.SP.800-60v1r1
41. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. In: Besold, T.R., Bordes, A., d'Avila Garcez, A.S., Wayne, G. (eds.) Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, CEUR Workshop Proceedings, vol. 1773, CEUR-WS.org (2016), URL https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf
42. Rao, R.S., Pais, A.R.: Jail-phish: An improved search engine based phishing detection system. Comput. Secur. **83**, 246–

267 (2019), https://doi.org/10.1016/J.COSE.2019.02.011, URL https://doi.org/10.1016/j.cose.2019.02.011
43. Rényi, A.: On measures of entropy and information. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics, vol. 4, pp. 547–562, University of California Press (1961)
44. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR **abs/1910.01108** (2019), URL http://arxiv.org/abs/1910.01108
45. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017, pp. 3–18, IEEE Computer Society (2017), https://doi.org/10.1109/SP.2017.41, URL https://doi.org/10.1109/SP.2017.41
46. Silvestri, F.: Mining query logs: Turning search usage data into knowledge. Found. Trends Inf. Retr. **4**(1-2), 1–174 (2010), https://doi.org/10.1561/1500000013, URL https://doi.org/10.1561/1500000013
47. Sousa, S., Kern, R.: How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. Artif. Intell. Rev. **56**(2), 1427–1492 (2023), https://doi.org/10.1007/S10462-022-10204-6, URL https://doi.org/10.1007/s10462-022-10204-6
48. Truex, S., Liu, L., Gursoy, M.E., Wei, W., Yu, L.: Effects of differential privacy and data skewness on membership inference vulnerability. In: First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, TPS-ISA 2019, Los Angeles, CA, USA, December 12-14, 2019, pp. 82–91, IEEE (2019), https://doi.org/10.1109/TPS-ISA48467.2019.00019, URL https://doi.org/10.1109/TPS-ISA48467.2019.00019
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 5998–6008 (2017)
50. Voorhees, E.M.: Overview of the TREC 2004 robust track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, NIST Special Publication, vol. 500-261, National Institute of Standards and Technology (NIST) (2004), URL http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf
51. Wagner, I., Eckhoff, D.: Technical privacy metrics: A systematic survey. ACM Comput. Surv. **51**(3), 57:1–57:38 (2018), https://doi.org/10.1145/3168389, URL https://doi.org/10.1145/3168389
52. Xu, Z., Aggarwal, A., Feyisetan, O., Teissier, N.: A differentially private text perturbation method using regularized mahalanobis metric. In: Proceedings

of the Second Workshop on Privacy in NLP, Association for Computational Linguistics (2020), https://doi.org/10.18653/v1/2020.privatenlp-1.2
53. Xu, Z., Aggarwal, A., Feyisetan, O., Teissier, N.: On a utilitarian approach to privacy preserving text generation. CoRR **abs/2104.11838** (Apr 2021), https://doi.org/10.48550/ARXIV.2104.11838
54. Yue, X., Du, M., Wang, T., Li, Y., Sun, H., Chow, S.S.M.: Differential privacy for text analytics via natural text sanitization. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 3853–3866, Association for Computational Linguistics, Online (Aug 2021), https://doi.org/10.18653/v1/2021.findings-acl.337, URL https://aclanthology.org/2021.findings-acl.337
55. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net (2020), URL https://openreview.net/forum?id=SkeHuCVFDr
56. Zhao, Y., Chen, J.: A survey on differential privacy for unstructured data content. ACM Comput. Surv. **54**(10s), 207:1–207:28 (2022), https://doi.org/10.1145/3490237, URL https://doi.org/10.1145/3490237
57. Zimmerman, S., Thorpe, A., Fox, C., Kruschwitz, U.: Investigating the interplay between searchers' privacy concerns and their search behavior. In: Piwowarski, B., Chevalier, M., Gaussier, É., Maarek, Y., Nie, J., Scholer, F. (eds.) Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, pp. 953–956, ACM (2019), https://doi.org/10.1145/3331184.3331280, URL https://doi.org/10.1145/3331184.3331280