Privacy-preserving Active Learning on Sensitive Data for User Intent Classification

Oluwaseyi Feyisetan
Amazon Research
sey@amazon.com

Thomas Drake
Amazon Research
draket@amazon.com

Borja Balle
Amazon Research
pigem@amazon.co.uk

Tom Diethe
Amazon Research
tdiethe@amazon.co.uk

Abstract

Active learning holds promise of significantly reducing data annotation costs while maintaining reasonable model performance. However, it requires sending data to annotators for labeling. This presents a possible privacy leak when the training set includes sensitive user data. In this paper, we describe an approach for carrying out privacy preserving active learning with quantifiable guarantees. We evaluate our approach by showing the tradeoff between privacy, utility and annotation budget on a binary classification task in a active learning setting.

Introduction

Preserving data privacy is an essential tenet required to maintain the bond of trust between consumers and corporations. Consumers expect their data to remain secure while being used to design better services for them without compromising their identities – especially while carrying out sensitive transactions and interactions. We define these potentially compromising and personally identifiable data as *sensitive data*. Annotated data drives the machine learning economy and sensitive data holds the key to building richer experiences for users interacting with modern AI interfaces. However, in a bid to get annotations, sensitive data in the wrong hands could lead to irreparable damage in terms of reputation and trust between data holders and their users.

This potential data transfer deserves greater monitoring in the era of human powered crowdsourcing and active learning. As niche classification tasks arise to power new applications, they often lack an abundance of pre-annotated datasets. With active learning, the learner can select a subset of the available data points to be annotated. This can exponentially reduce (Settles 2010) (in some cases) the number of training queries required. However, the cost (Dasgupta 2011; Arora, Nyberg, and Rosé 2009; Settles 2010) of labelling machine learning datasets is traditionally viewed as a function of the expert, time or price.

In this paper, we argue that, for non-public datasets, the cost of learning the true labels should also factor in the privacy of the information contributed by the data owners to the data custodians. As a result, the active learning condition (beyond simply selecting the best examples) becomes two-fold when submitting a *batch* of data for annotation: (1) labeling this selected subset leads to the greatest increase in our machine learning model performance, (2) the probability of revealing any query that can uniquely identify a specific user is very small (and quantifiable by a privacy parameter).

Contributions Recent studies into privacy and machine learning have focused on preserving model parameters from leaking training data (Papernot et al. 2016; Hamm, Cao, and Belkin 2016). See (Ji, Lipton, and Elkan 2014) for a recent survey. However, in this paper, we address the privacy preserving requirement from the point of view of training samples that are sent to annotators from an active learning model. To the best of our knowledge, this is the first paper that views preserving privacy in machine learning from this angle. We also describe how techniques such as k-anonymity does not provide sufficient privacy guarantees and how this can be improved using differential privacy (DP). We describe how to do this by providing experimental results after discussing an approach that leverages one of the DP algorithms from literature.

Background

In this section, we present an introduction to active learning and the privacy challenge of outsourcing queries to the crowd. We then describe k-anonymity, its shortcoming in providing an adequate privacy model for active learning and how this can be improved with differential privacy.

Active Learning

The central premise of active learning is that a model is able to perform as well with less data, if a learner can select the training examples that provide the highest information (Settles 2010). Formally described, using a classification task: let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ where the goal is to output a label $\mathcal{Y}$ from the label space $\{\pm 1\}$ given an input from the feature space $\mathcal{X}$. The learner receives a *batch* of i.i.d. draws (x_1, y_1), ..., (x_n, y_n) over the unknown underlying distribution $\mathcal{D}$. The value of y_i is unknown unless an annotation request is made by the learner. The objective is to select a hypothesis function $h : \mathcal{X} \rightarrow \mathcal{Y}$, where err($h$) = $\mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$ is small. Given that $\mathcal{H}$ is the space of

Accepted at Privacy-Enhancing Artificial Intelligence and Language Technologies. PAL 2019, March 2019, Palo Alto, CA USA

all hypothesis, and $h* = \text{argmin}\{\text{err}(h) : h \in \mathcal{H}\}$ is the hypothesis with minimum error, the aim of active learning is to select a hypothesis $h \in \mathcal{H}$ with error err(h) within reasonable bounds of err($h*$) by using few annotation requests (i.e., few compared to a passive learner).

Various strategies have been proposed to implement an active learner. One is *uncertainty sampling* (Lewis and Gale 1994) which attempts to select the query x' that the model is least convinced about; i.e., $x' = \text{argmax}_x\, 1 - P_\theta(\hat{y}|x)$, where $\hat{y}$ is the label with the highest posterior for model θ and x is maximized over the range of all the unlabeled examples in the training pool. Other approaches to uncertainty sampling use either the *margin* between the two most probable classes $\hat{y}_1$ and $\hat{y}_2$; i.e., $x' = \text{argmin}_x\, P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$ or a general entropy-based uncertainty over all the possible $\hat{y}_i$ classes; i.e., $x' = \text{argmax}_x - \sum_i P_\theta(\hat{y}_i|x) log P_\theta(\hat{y}_i|x)$.

The main privacy issue with active learning stems from the need to scale the annotation process by crowdsourcing the labels via an open call (Howe 2006). Whenever you make a request to an external resource, you pay a privacy cost by transmitting the information to be annotated. This problem is compounded when there is only one oracle (Avidan and Butman 2007) or collusion among crowd workers. In this paper, we describe privacy notions that can be used to address these concerns along the privacy-utility tradeoff spectrum.

Privacy-preserving machine learning

***k*-Anonymity** At first glance, a straightforward approach for addressing the privacy concerns of active learning could be through k-anonymity (Sweeney 2002; Di Castro et al. 2016); i.e., ensuring each query that is sent out for crowdsourcing occurs at least k times. In deploying k-anonymity, the first step involves identifying a set of *quasi-identifiers*. In our context, these are user queries which can be potentially combined with an externally available dataset to uniquely identify a user. The *frequency set* of these quasi-identifiers represent the number of occurrences in the dataset. We therefore say that a dataset satisfies k-anonymity relative to the quasi-identifiers if when it is projected on an external dataset, the frequency set occurs greater than or equal to k times.

To achieve k-anonymity when the size of the frequency set is less than a desired k, the attributes are anonymized by either generalizing or suppressing the information. For example, marital status attributes listed as *married, divorced or widowed* are generalized as *once married*, while the ethnicity is redacted as *****.

Despite its promise, k-anonymity has fundamental challenges, some of which are exacerbated by our unstructured data domain. First, (Aggarwal 2005) demonstrated that k-anonymity suffers from the *curse of dimensionality* since generalization (such as with traditional database columns), requires co-occurrence of words across different examples, but unstructured data such as text phrases tend to follow a heavy-tailed distribution that have a low co-occurrence of words. Secondly, the choice of quasi-identifiers might exclude the selection of some useful sensitive attributes which could then be used for re-identification attacks. This led to other approaches such as l-diversity (Machanavajjhala et al. 2006) and t-closeness (Li, Li, and Venkatasubramanian 2007) to handle sensitive attributes. In our implementation, we subsume the quasi-identifiers to include the entire user query.

Therefore, by 'hiding in the crowd' of k, a user has received some assurance from k-anonymity that their sensitive query will not be outsourced from the active learning model unless it passes a meaningful threshold. However, stronger formal privacy guarantees are required to demonstrate that given the user's query, an attacker cannot decide where it came from with certainty. With k-anonymity, we are unable to directly quantify a privacy loss value, nor state the bounds of the guarantee of this loss. These two quantities are obtainable from a differential privacy model which we now describe.

Differential Privacy To motivate our discourse on why we need stronger privacy guarantees than what k-anonymity provides, we consider a hypothetical scenario: Would a user be comfortable asking an AI agent a sensitive question, with the knowledge that the question will be *possibly* used to further train agent's learning model? We denote the training data available to the model before the user submission as $\mathcal{D}$, and the data after the user question as $\mathcal{D}'$. These are *adjacent datasets* differing on only one record. We posit that a user c will be comfortable if (1) $\mathcal{Q}(\mathcal{D}) = \mathcal{Q}(\mathcal{D}')$ where $\mathcal{Q}$ is a query over the dataset; and (2) $P(\mathcal{S}(c)|D') = P(\mathcal{S}(c))$ where $\mathcal{S}$ is a user secret. These 2 points are articulated in *Dalenius's Desideratum* (Dwork 2011) that:

> Anything that can be learned about a respondent from the statistical database should be learnable without access to the database

However, we can't make these exact guarantees because datasets are meant to convey information and they will have no utility if these points were true.

What Differential Privacy (Dwork 2011; Dwork and Roth 2014) offers is a strong privacy guarantee on adjacent datasets (taking our AI agent example), that: the example selected for active learning will be very similar whether or not the user added their sensitive question. This means, an adversarial annotator receiving a random training query cannot guess with certainty if the query was from dataset $\mathcal{D}$ (which doesn't include the user's query) or $\mathcal{D}'$ (which includes it).

With this, we state that, a randomized algorithm $\mathcal{M} : \mathbb{N}^{\mathcal{X}} \to \mathcal{C}$ that receives as input a dataset $\mathcal{D}$ with records from a universe $\mathcal{X}$ and outputs an element from $\mathcal{C}$ is (ε, δ)-differentially private if for every pair of databases $\mathcal{D}$ and $\mathcal{D}'$ differing in one record and every possible set of outputs $C \subseteq \mathcal{C}$ we have

$$\mathsf{Pr}[\mathcal{M}(\mathcal{D}) \in C] \leq e^{\varepsilon}\mathsf{Pr}[\mathcal{M}(\mathcal{D}') \in C] + \delta \tag{1}$$

The δ parameter accounts for a $< 1/||\mathcal{D}||_1$ relaxed chance of the ϵ guarantee not holding – otherwise, it will be equivalent to just selecting a random sample on the order of the size of the dataset. One benefit of the differential privacy model is that it has a quantifiable, non binary value for *privacy loss* which helps in deciding to comparatively select one algorithm over the other. We observe an output of the random algorithm $C \sim \mathcal{M}(D)$ where we believe that C was more likely produced by $\mathcal{D}$ and not $\mathcal{D}'$, then the privacy loss from the query that yields C on an auxiliary input x is:

$$\mathcal{L}(C; \mathcal{M}, x, \mathcal{D}, \mathcal{D}') \overset{\text{def}}{=} \ln(\frac{Pr[\mathcal{M}(\mathcal{D}) = C]}{Pr[\mathcal{M}(\mathcal{D}') = C]}) \qquad (2)$$

So we surmise that differential privacy promises to prevent a user from sustaining *additional* damage by including their data in a dataset; and the privacy loss obtained is ϵ with probability $\geq 1 - \delta$.

A common method for making the results of a statistical query differentially private involves adding Laplacian noise proportional to either the query's global sensitivity (Dwork 2008; Dwork et al. 2006) or the smooth bound of the local sensitivity (Nissim, Raskhodnikova, and Smith 2007) (where sensitivity $\Delta f = \max ||f\mathcal{D} - f\mathcal{D}'||$). However, for non-continuous domains, adding noise can result in unintended consequences that completely wipe out the utility of the results e.g., (Dwork and Roth 2014) describe how attempting to add noise to the query for the optimal price for an auction could drive the revenue to zero.

Research has however shown that apart from providing reasonable and well understood protection from inadvertent exposure (Di Castro et al. 2016), k-anonymity can also be used as a launchpad for achieving quantifiable differential privacy without the utility loss that comes from applying noise (Li, Qardaji, and Su 2012; Soria-Comas et al. 2014).

Privacy Preserving Active Learning Framework

This section introduces our proposed framework for carrying out active learning with privacy guarantees on queries that are sent to an external oracle. It presents the task we try our approach on, highlights the considerations that drive our choices and lays out a high level pseudo-code of our approach.

Task model

Our task consists of a very large dataset of user queries $\mathcal{U} = \{u_1, ..., u_n\}$ that represent the user intent (we map the queries to the intent and do not extract specific quasi-identifiers in order to prevent leakages from un-captured sensitive attributes). Our pipeline consists of an active learning model which learns a binary classifier, predicting if a user intent belongs to a specified class or not. The model is bootstrapped with a golden set of user queries and their associated intents. Subsequent queries from a fixed pool are added to a RANKEDEXAMPLEPOOL where they are ordered by confidence/uncertainty (Gal and Ghahramani 2016) from our deep learning model.

To train the model, it first draws on the golden set, then we make a $next\_example$ call to draw an uncertain query from the pool with the criteria that knowing the accurate intent of this query gives the best performance increase to the model while preserving privacy. The query is then outsourced to external annotators and the annotated labels are re-incorporated into the model training process.

Considerations

Given the size and projected scale of our dataset ($\approx 10^9$ queries), we decide to employ randomized probabilistic algorithms in estimating if a query satisfies k-anonymity. Compute and memory resources are thus freed up for training and retraining the model rather than maintaining the frequency and cardinality of incoming queries. Each algorithm (detailed below) is adjusted to prevent over-estimations which could erode the privacy guarantees. Furthermore, after a query is presumed to satisfy k-anonymity, only 1 of the k queries is sent to n external annotators to prevent an aggregation of privacy losses.

Approach

In this paper, we adopt the differential privacy algorithm from (Li, Qardaji, and Su 2012) but we utilize it in an active learning setting to select a subset of training examples to send for crowdsourcing. We also note that other DP methods that have been designed for search logs and include a form of k parameter aggregation such as: (Korolova et al. 2009), ZEALOUS (Gotz et al. 2012) and SafeLog (Zhang et al. 2016) can be implemented to obtain similar results.

We take a two-stepped approach to extend k-anonymity to yield a quantifiable differentially private active learning model taking a cue from how (Li, Qardaji, and Su 2012) demonstrated the use of pre-sampling to achieve differential privacy with k-anonymity. This is predicated on THEOREM 1 from (Li, Qardaji, and Su 2012) which states that: given an algorithm $\mathcal{M}$ which satisfies $(\beta_1, \epsilon_1, \delta_1)$-differential privacy under sampling, then $\mathcal{M}$ also satisfies $(\beta_2, \epsilon_2, \delta_2)$-differential privacy under sampling for any $\beta_2 < \beta_1$ where

$$\epsilon_2 = \ln\left(1 + \left(\frac{\beta_2}{\beta_1}(e^{\epsilon_1} - 1)\right)\right); \delta_2 = \frac{\beta_2}{\beta_1}\delta_1 \qquad (3)$$

Therefore, k-anonymity on our full dataset (i.e., $\beta_1 = 1$) can instead be preceded by a mechanism that samples each row of its input with probability β_2, with k-anonymity then applied to the resulting sub-sample to yield ϵ_2, δ_2-differential privacy for $\epsilon_2 = \ln(1 + (\beta_2(e^{\epsilon_1} - 1)))$ within the bounds $\delta_2 = \beta_2\delta_1$. Thus the effect of sampling serves to amplify pre-existing privacy guarantees (Balle, Barthe, and Gaboardi 2018).

Furthermore, we harden our k-anonymity to offer 'safe' k-anonymization by aggregating the queries by frequency rather than using a distance based measure (LeFevre, DeWitt, and Ramakrishnan 2006). The benefit we get from this is that no query within our set of k contains any extraneous sensitive text which could be used as a source of re-identification or to carry out reconstruction attacks.

The next sections describe: how we carry out our sampling to ensure we select useful candidates in an efficient manner, and how we estimate k-anonymity using the queries.

Efficient sub-sampling for active learning

Given a multiset of query sets $\mathcal{M} = \{U_1, ..., U_s\}$ with repetitions where a given U_i is a tuple $\langle u_i, ..., u_k \rangle$, and a sampling rate β, our objective is to return a sub-sample from which to

Algorithm 1: Privacy-preserving Active Learning

```
Let $\beta$ be the sampling rate // $\beta = \beta_2$ from (3)
Let $k$ be the anonymity parameter
Let $l$ be number of samples for variance calculation
Data: Input multiset of samples $x$ with unknown labels
    $y$ as: $\mathcal{U} = \{x_1, y_1\}, ..., \{x_n, y_n\}$
Result: $\mathcal{U}'$ filtered private multiset

Bootstrap AL probablistic model $P_\theta$ with labeled
  utterances $\mathcal{L}$
Retrieve random sub-sample $\mathcal{U}'$ of size $\beta n$
Pool creation: to add queries to the
  RankedExamplePool
for $\{x, y\} \in \mathcal{U}'$ do
    retrieve freq(x);
    if freq(x) $\geq k$ then
        retrieve variance $\phi_x = var(P_\theta(\hat{y}|x) : 1..l)$ on $u$;
            // computed using $l$ draws
        add $\langle x, \phi_x \rangle$ to RankedExamplePool;
    else
        remove $\{x, y\}$ from $\mathcal{U}'$;
    end
end

Acquire labels for top examples sorted by $\phi_x$
  descending as $\{\hat{y}_1, ..., \hat{y}_{\hat{n}}\}$
Set $\mathcal{U}'$ to be $(x_1, \hat{y}_1), ..., (x_{\hat{n}}, \hat{y}_{\hat{n}})$
Update model: draw the next training example from $\mathcal{U}'$
  over $\phi_x$
    get $\hat{x} = \text{argmax}_u\, 1 - P_\theta(\hat{y}|x)$; - # sample we are
  least confident of
    retrain learning model using $\{\hat{x}, \hat{y}\}$

return $\mathcal{U}'$
```

carry out k-anonymization before training our active learner. Let *n* be the number of distinct query sets —$\{U_1, ..., U_s\}$— with elements $\{e_1, ..., e_n\}$. For a very large dataset size *s*, we seek to estimate $\hat{n}$ using only *m* registers where $m << n$. The number of distinct queries in our sample set therefore become $\beta\hat{n}$.

To estimate the cardinality $\hat{n}$, we utlize the HyperLogLog algorithm by (Flajolet et al. 2007). HyperLogLog is a probabilistic cardinality estimator that uses a very small memory footprint (≈ 12kb per key) for a low standard error (≈ 0.81%) while scaling up to dataset sizes as large as 2^{64} items[1].

For each incoming $U_i : \langle u_i, ..., u_k \rangle$, a hash $h(U_i)$ is computed and converted to base 2. The *b* least significant bits are used to identify the register location to modify, where $2^b = m$ or $\log_2 m$. With the remaining bits *w*, a count *p(w)* is made of the number of running 0*s* up to the leftmost 1. For a very large, uniformly distributed multiset of random numbers, 2 raised to the maximum value of *p(w)* gives a wide approximate of the cardinality. To correct this, HyperLogLog breaks the multiset into subsets and uses the harmonic mean of the subsets.

After determining our sample size, the next step is to draw a random set of unique samples without replacement up to $\beta\hat{n}$. We keep each element in the dataset with probability β. The ensuing sub-sample represents the new dataset in our RankedExamplePool from which we will carry out our k-anonymization.

Estimating *k*-anonymity using query frequency

Given a multiset of query sets $\mathcal{M} = \{U_1, ..., U_s\}$ with repetitions such that the frequency of U_i is f_{U_i} and U_i is a tuple $\langle u_i, ..., u_k \rangle$. For a very large dataset size s, we seek to estimate $\hat{f_{U_i}}$ using sub-linear space. To estimate the query frequency, we use the Count-Mean-Min with conservative update (Goyal, Daumé III, and Cormode 2012) sketch algorithm which is an improvement on the proposed Count-Min sketch algorithm by (Cormode and Muthukrishnan 2005). For each incoming $U_i : \langle q_i, ..., q_k \rangle$, d different hashes of the queries is computed and a counter indexed by each hashed result is incremented. To return the frequency, the minimum over all d index locations for Q_i is returned. To further reduce the potential of error from over-estimation, conservative updates are employed to increment only the minimum counter from the d indexes, and an estimated noise is further deducted from the result.

Therefore after initial pre-sampling step, we select only queries which occur at least k times. These queries are then added to the RankedExamplePool where the *next_example* is drawn based on the element with the highest uncertainty measure. The benefit of using the frequency to satisfy k-anonymity rather than using partitioning, clustering and recoding, or distance based algorithms, is to prevent attacks that rise from an attackers a-priori knowledge of a dataset. For example, a cluster of k with one sensitive or extreme outlier (e.g., a cluster of incomes within zip code with one UHNW outlier becomes easily identifiable by an attacker even though the aggregation was based on nearest neighbors).

Experiments

Our work seeks to demonstrate quantifiable privacy preserving guarantees in an active learning setting by taking a pre-sampling approach before carrying out k-anonymization. We evaluate our approach on an internal dataset used for intent classification on voice devices.

Datasets

The Intent Classifier dataset consists of a subset of queries from February 2018. The dataset is used to train a model which determines a binary intent for a user. The dataset consists of 2.5M queries comprising 58K distinct data points. Each record contains a user query and a label indicating if it is categorized as a Positive or Negative intent query. Part of the dataset has also been previously discussed and described by (Yang et al. 2018). Figures 1c and 1d show the

[1] Values taken for the Redis implementation of HyperLogLog - http://antirez.com/news/75

nature of the dataset with a histogram and plot of the frequency distribution of the queries. As expected with textual data, there is a long tail of queries which were observed just once (making up $\approx 60\%$ of the dataset). The dataset consists of 63% of queries labelled as POSITIVE intents vs 27% being NEGATIVE.

Experiment setup

The experiment task was binary intent classification in an active learning setting. We created a new baseline model which predicts POSITIVE and NEGATIVE intents. For the experiments, the model was initially bootstrapped with $1,000$ labeled examples. The active learner then queries a data pool to get a batch of additional training examples to improve the model. The active learning strategy was uncertainty sampling based on confidence scores.

The confidence and uncertainty scores for the active learning model were obtained from a Bayesian deep learning model described in (Yang et al. 2018) where model uncertainty, quantified by Shannon entropy is $\mathcal{U}(x) = -\sum_c(\frac{1}{T}\sum_t \hat{p}_c)\log(\frac{1}{T}\sum_t \hat{p}_c)$ and $\frac{1}{T}\sum_t \hat{p}_c$ is the averaged predicted probability of class c for x, sampled T times by Monte Carlo dropout. A histogram of the confidence and uncertainty scores can be seen in Figures 1a and 1b.

We simulated the probability of the crowd annotators returning the correct answers to the requested queries by drawing from a normal distribution with mean centered at 0.65 and standard deviation 0.01 (see (Yang et al. 2018)'s Figure 2(a) for more).

Evaluation metrics

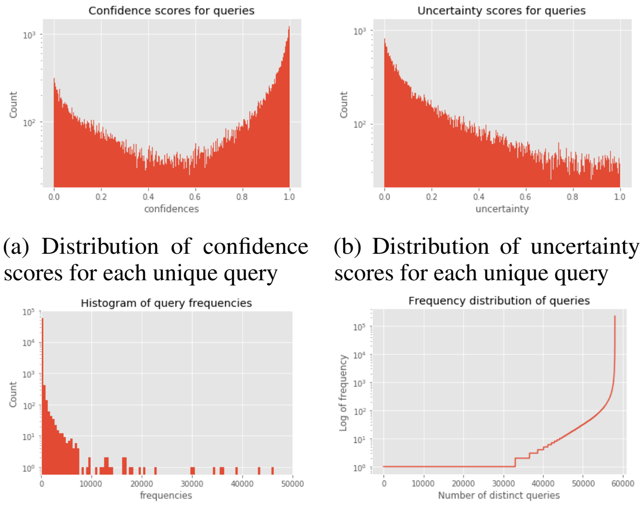
To evaluate our results, we compared the annotation accuracy between the baseline model, and the models trained with active learning and our privacy preserving model. We vary the sub-sampling parameter β and the anonymization factor k while training our model and recording its accuracy. We set the evaluation data at $5,000$ samples (i.e., about 10% of the dataset). We also provide privacy guarantee values from numerical computations of ϵ and δ and highlight in the appendix, what values of k and β provide those levels of guarantees.

Baseline condition train standard classification model. Sub-sampling parameter $\beta = 1$, anonymization factor $k = 1$ i.e., using the entire dataset

Experiment conditions train classification model using privacy preserving active learning. Sub-sampling parameter varied at $\beta = \{0.1, 0.3, 0.6, 0.9\}$, anonymization factor varied at $k = \{1, 20, 100, 200, 500\}$

Results

The results of our experiments are presented in Figures 2, 3 and 4. Our findings provide insight to the tradeoffs between privacy, utility and our annotation budget.



(a) Distribution of confidence scores for each unique query

(b) Distribution of uncertainty scores for each unique query

(c) Histogram of frequency distribution

(d) Log frequency distribution of queries

Figure 1: A view into the training dataset

Privacy vs Utility Tradeoff

Figure 2 highlights the privacy–utility tradeoff which occurs as a result of varying β and k. As expected, as the value of k, gets smaller, i.e., by selecting more items in the tail of the dataset, we are able to improve the accuracy of our model. This however has the effect of degrading our privacy guarantees. Similarly, by providing privacy amplification by subsampling, the utility of our model suffers. Figure 2 paints a wholistic picture of this by showing how by tuning the values of β and k, we can arrive at the same values of accuracy.

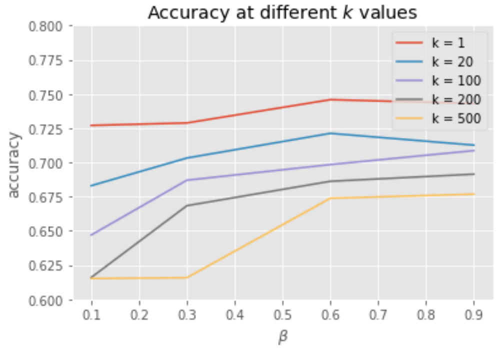


Figure 2: Accuracy at different k values

Annotation budget

Figure 3 describes how our annotation budget changes for different privacy settings. With a stronger privacy model, we incur less cost as a function of less annotation requests. By reading across the graph, we also discover that the same budget can be realized from different privacy configurations: e.g., by subsampling with $\beta = 0.1$ and selecting $k = 20$, we incur the same budget as $\beta = 0.3$ and $k = 100$ and therefore, the same accuracy (from Figure 2 above).

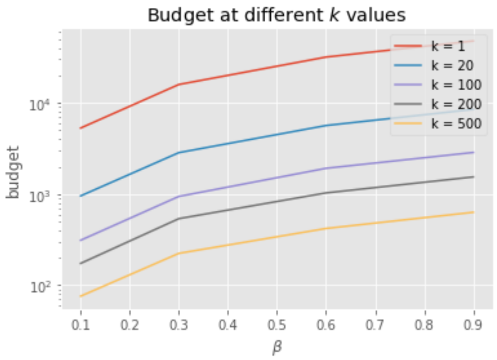


Figure 3: Budget at different k values

Budget vs Accuracy

We established from Figure 2 and Figure 3, the relationship between privacy and accuracy, and between privacy and our annotation budget. Since we can obtain the same level of accuracy and budget requirements from different parameter values, Figure 4 highlights how an increase in budget affects our overall model accuracy. Increasing the budget initially accelerates the improvement of our model, however, the utility gains quickly slow down. For example, after $30,000$ labels, we do not see any significant increase in model accuracy.

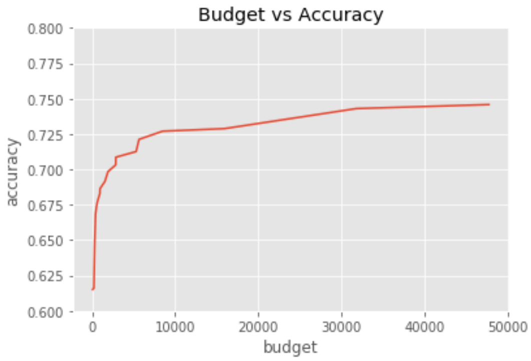


Figure 4: Budget vs Accuracy

These results can serve as a guideline in selecting appropriate privacy parameters for different annotation budgets in a way that is more representative of the dataset. For example, for a fixed annotation budget, you can reduce β to select more data points from the tail of the dataset (i.e., a smaller k). This variation can also be done by starting with a target accuracy score and varying β and k. The results also demonstrate that by sacrificing some utility gains, we can make stronger privacy guarantees and reduce our annotation budget when carrying out active learning.

Conclusion

We now briefly revisit our results in the light of our hypothesis. We also discuss the limitations of our process, its implication to the broader discourse on privacy and machine learning and conclude with future work.

We apply the approach from (Li, Qardaji, and Su 2012) to offer privacy guarantees when training models with active learning which requires sending unlabelled examples to an external oracle. Our results join the conversation on differential privacy and machine learning (Ji, Lipton, and Elkan 2014) with particular reference to preserving the privacy of users.

Our results show that by taking a small performance hit, we can achieve similar accuracy scores with a smaller annotation budget and stronger privacy guarantee. One limitation however is that we have only reported results on a binary classification task. We are currently expanding our approach by designing a new algorithm for differential privacy on text. We show that the accuracy loss increases as task complexity increases. Therefore, if we were to apply the approach in this work to other NLP tasks, e.g. multi-class classification or question answering, we will expect the accuracy loss to be greater.

Another limitation of our approach and potentially other $k-$parameter based approaches ((Korolova et al. 2009),(Gotz et al. 2012),(Zhang et al. 2016)) to differential privacy for text is that it will not work for tasks where almost all the data is unique i.e., k is essentially 1 (e.g., a datasets of emails or movie reviews). Therefore, a different approach is needed to provide quantifiable privacy guarantees without resorting to k-anonymity.

We believe that this is an area worthy of further research in order to further quantify the true cost of privacy in crowdsourcing and machine learning. We have already begun further work to address two of the limitations reported in this current paper.

Appendix

Table 1 lays out a grid of ϵ, δ scores and the corresponding sampling parameter β and anonymization factor k required to satisfy that level of (ϵ,δ)-differential privacy. The shaded region presents a $\beta \times k$ high level view of how to achieve a desired level of privacy.

A few insights can be gleaned from the results, the most obvious being that strong privacy requirements, (indicated by small ϵ and δ scores as we traverse the table towards the bottom left corner), require a higher anonymization factor k and smaller sampling rate β. This is also observed by the fact that lowering the k factor only preserves privacy at the highest displayed ϵ value of 1.0 and $\delta = 1 \times 10^{-6}$ (at the top right corner of the table).

Observing β individually, we also note that, when k and ϵ are at fixed values, as δ decreases, i.e., to obtain stronger privacy guarantees, we need to lower the sampling rate β. This indicates as β decreases, privacy guarantees increase. Similarly, observing k by fixing the values of β and ϵ, demonstrates that increasing k improves our privacy guarantees.

Table 1: Selected values of ϵ against δ: the shaded regions show the accuracy scores from our experiments and what $\beta = \{0.1, 0.3, 0.6, 0.9\}$ vs $k = \{20, 100, 200, 500\}$ values provide the (ϵ,δ)-differential privacy guarantee

δ		ϵ															
		0.25				0.5				0.75				1.0			
		$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.6$	$\beta = 0.9$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.6$	$\beta = 0.9$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.6$	$\beta = 0.9$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.6$	$\beta = 0.9$
1×10^{-6}	k = 20					68.3				68.3				68.3	70.3		
	k = 100	64.7	68.7			64.7	68.7			64.7	68.7	69.8		64.7	68.7	69.8	
	k = 200	61.6	66.8			61.6	66.8	68.6		61.6	66.8	68.6		61.6	66.8	68.6	
	k = 500	61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3	67.6
1×10^{-9}	k = 20					68.3				68.3				68.3			
	k = 100	64.7				64.7	68.7			64.7	68.7			64.7	68.7	69.8	
	k = 200	61.6	66.8			61.6	66.8			61.6	66.8	68.6		61.6	66.8	68.6	
	k = 500	61.5	61.6			61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3	
1×10^{-12}	k = 20									68.3				68.3			
	k = 100	64.7				64.7	68.7			64.7	68.7			64.7	68.7		
	k = 200	61.6				61.6	66.8			61.6	66.8	68.6		61.6	66.8	68.6	
	k = 500	61.5	61.6			61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3	
1×10^{-15}	k = 20																
	k = 100	64.7				64.7				64.7	68.7			64.7	68.7		
	k = 200	61.6				61.6	66.8			61.6	66.8			61.6	66.8	68.6	
	k = 500	61.5	61.6			61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3	

References

[Aggarwal 2005] Aggarwal, C. C. 2005. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, 901–909. VLDB Endowment.

[Arora, Nyberg, and Rosé 2009] Arora, S.; Nyberg, E.; and Rosé, C. P. 2009. Estimating annotation cost for active learning in a multi-annotator environment. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 18–26. Association for Computational Linguistics.

[Avidan and Butman 2007] Avidan, S., and Butman, M. 2007. Efficient methods for privacy preserving face detection. In *Advances in neural information processing systems*, 57–64.

[Balle, Barthe, and Gaboardi 2018] Balle, B.; Barthe, G.; and Gaboardi, M. 2018. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *arXiv preprint arXiv:1807.01647*.

[Cormode and Muthukrishnan 2005] Cormode, G., and Muthukrishnan, S. 2005. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms* 55(1):58–75.

[Dasgupta 2011] Dasgupta, S. 2011. Two faces of active learning. *Theoretical computer science* 412(19):1767–1781.

[Di Castro et al. 2016] Di Castro, D.; Lewin-Eytan, L.; Maarek, Y.; Wolff, R.; and Zohar, E. 2016. Enforcing k-anonymity in web mail auditing. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 327–336. ACM.

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

[Dwork et al. 2006] Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 265–284. Springer.

[Dwork 2008] Dwork, C. 2008. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, 1–19. Springer.

[Dwork 2011] Dwork, C. 2011. A firm foundation for private data analysis. *Communications of the ACM* 54(1):86–95.

[Flajolet et al. 2007] Flajolet, P.; Fusy, É.; Gandouet, O.; and Meunier, F. 2007. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *AofA: Analysis of Algorithms*, 137–156. Discrete Mathematics and Theoretical Computer Science.

[Gal and Ghahramani 2016] Gal, Y., and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059.

[Gotz et al. 2012] Gotz, M.; Machanavajjhala, A.; Wang, G.; Xiao, X.; and Gehrke, J. 2012. Publishing search logs – a comparative study of privacy guarantees. *IEEE Transactions on Knowledge and Data Engineering* 24(3):520–532.

[Goyal, Daumé III, and Cormode 2012] Goyal, A.; Daumé III, H.; and Cormode, G. 2012. Sketch algorithms for estimating point queries in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1093–1103. Association for Computational Linguistics.

[Hamm, Cao, and Belkin 2016] Hamm, J.; Cao, Y.; and Belkin, M. 2016. Learning privately from multiparty data. In *International Conference on Machine Learning*, 555–563.

[Howe 2006] Howe, J. 2006. The rise of crowdsourcing. *Wired magazine* 14(6):1–4.

[Ji, Lipton, and Elkan 2014] Ji, Z.; Lipton, Z. C.; and Elkan, C. 2014. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*.

[Korolova et al. 2009] Korolova, A.; Kenthapadi, K.; Mishra, N.; and Ntoulas, A. 2009. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web*, 171–180. ACM.

[LeFevre, DeWitt, and Ramakrishnan 2006] LeFevre, K.; DeWitt, D. J.; and Ramakrishnan, R. 2006. Mondrian multidimensional k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, 25–25. IEEE.

[Lewis and Gale 1994] Lewis, D. D., and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 3–12. Springer-Verlag New York, Inc.

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

[Li, Qardaji, and Su 2012] Li, N.; Qardaji, W.; and Su, D. 2012. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, 32–33. ACM.

[Machanavajjhala et al. 2006] Machanavajjhala, A.; Gehrke, J.; Kifer, D.; and Venkitasubramaniam, M. 2006. l-diversity: Privacy beyond k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, 24–24. IEEE.

[Nissim, Raskhodnikova, and Smith 2007] Nissim, K.; Raskhodnikova, S.; and Smith, A. 2007. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 75–84. ACM.

[Papernot et al. 2016] Papernot, N.; Abadi, M.; Erlingsson, U.; Goodfellow, I.; and Talwar, K. 2016. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*.

[Settles 2010] Settles, B. 2010. Active learning literature survey. 2010. *Computer Sciences Technical Report* 1648.

[Soria-Comas et al. 2014] Soria-Comas, J.; Domingo-Ferrer, J.; Sánchez, D.; and Martínez, S. 2014. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal* 23(5):771–794.

[Sweeney 2002] Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05):557–570.

[Yang et al. 2018] Yang, J.; Drake, T.; Damianou, A.; and Maarek, Y. 2018. Leveraging crowdsourcing data for deep active learning an application: Learning intents in Alexa. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 23–32. International World Wide Web Conferences Steering Committee.

[Zhang et al. 2016] Zhang, S.; Yang, G. H.; Singh, L.; and Xiong, L. 2016. Safelog: Supporting web search and mining by differentially-private query logs. In *2016 AAAI Fall Symposium Series*.

Privacy-preserving Active Learning on Sensitive Data for User Intent Classification

Oluwaseyi Feyisetan, Thomas Drake, Borja Balle, Tom Diethe

Amazon Research

INTRODUCTION

Active Learning [1] holds promise. But, it requires sending data to annotators for labeling. This presents a possible privacy leak when the training set includes sensitive data.

We want to answer the question: which of these queries should be manually annotated and added to the ML model training data while satisfying two conditions:

1. we do not reveal any compromising or personally identifiable data to annotators, and
2. labeling this selected subset leads to the greatest increase in model performance.

PRIVACY VIA RANDOM SAMPLING & PRUNING

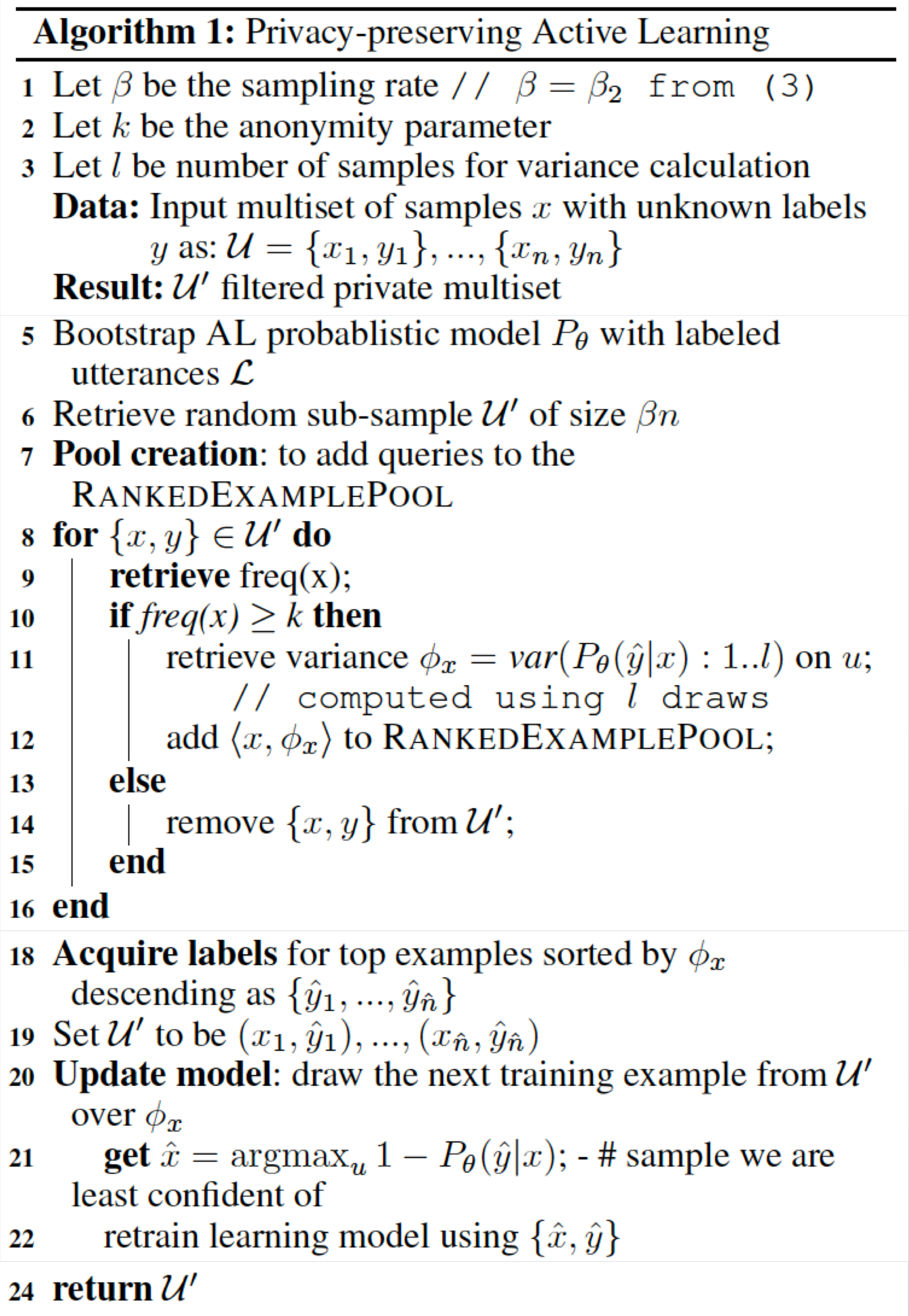
A randomized algorithm M is (ε, δ)-differentially private [2] if for every pair of databases D and D′ differing in one record and every possible set of outputs *C* ⊆ C we have:

$$\Pr[\mathcal{M}(\mathcal{D}) \in C] \leq e^{\epsilon}\Pr[\mathcal{M}(\mathcal{D}') \in C] + \delta$$

By preceding k-anonymity with a randomized sampling step with parameter β, [3] introduces (β, ε, δ) differential privacy under sampling. The parameters are linked as:

$$d(k,\beta,\epsilon) \quad = \max_{n:n\geq\left\lceil \frac{k}{\gamma}-1\right\rceil} \sum_{j>\gamma n}^{n} f(j;n,\beta), \qquad \textit{where } \gamma = \frac{(e^{\epsilon}-1+\beta)}{e^{\epsilon}}. \quad \delta = d(k,\beta,\epsilon)$$

PRIVACY PRESERVING ACTIVE LEARNING

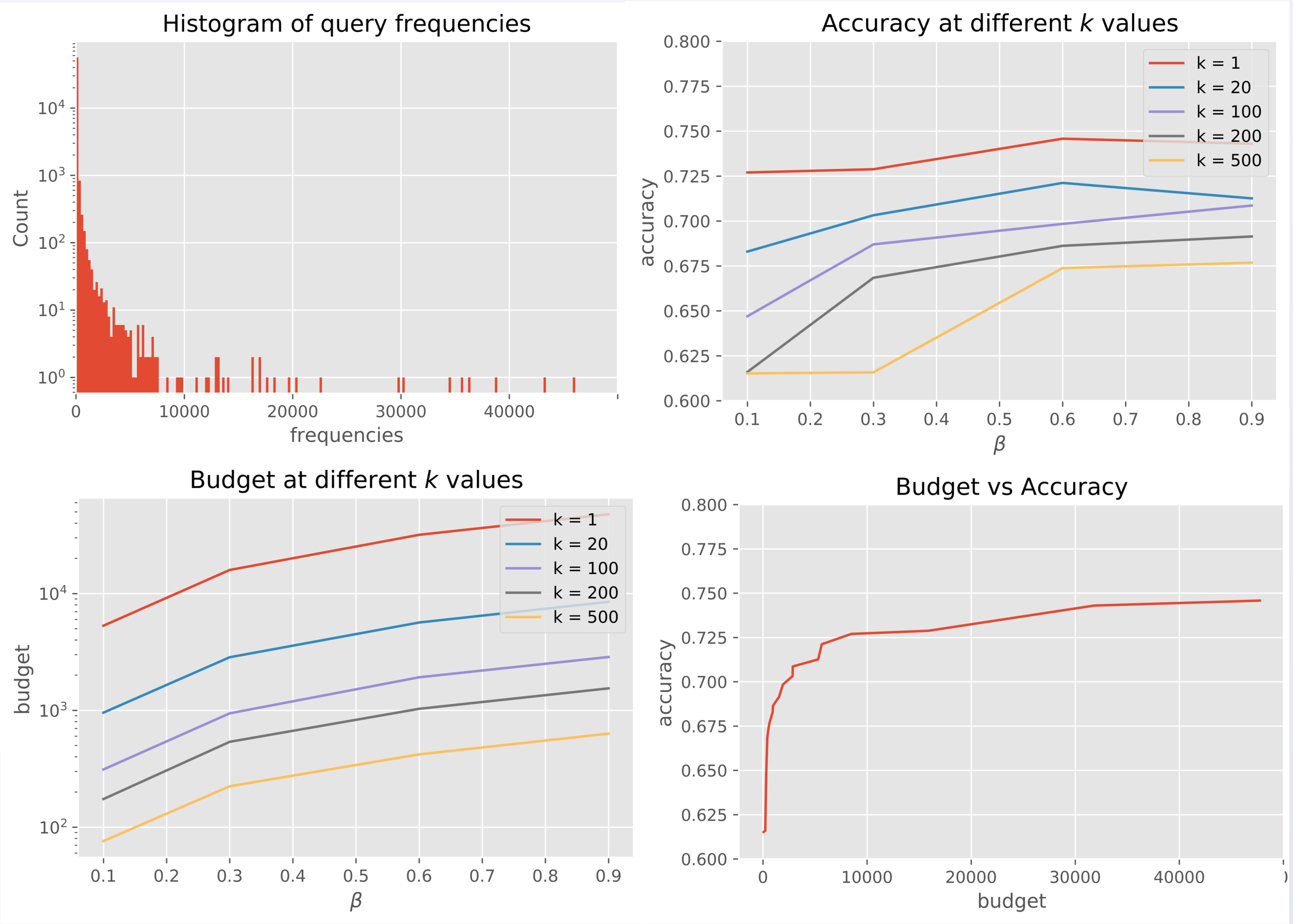


Algorithm 1: Privacy-preserving Active Learning

```
Let $\beta$ be the sampling rate // $\beta = \beta_2$ from (3)
Let $k$ be the anonymity parameter
Let $l$ be number of samples for variance calculation
Data: Input multiset of samples $x$ with unknown labels
        $y$ as: $\mathcal{U} = \{x_1, y_1\}, ..., \{x_n, y_n\}$
Result: $\mathcal{U}'$ filtered private multiset
Bootstrap AL probablistic model $P_\theta$ with labeled
  utterances $\mathcal{L}$
Retrieve random sub-sample $\mathcal{U}'$ of size $\beta n$
Pool creation: to add queries to the
  RankedExamplePool
for $\{x, y\} \in \mathcal{U}'$ do
    retrieve freq(x);
    if $freq(x) \geq k$ then
        retrieve variance $\phi_x = var(P_\theta(\hat{y}|x) : 1..l)$ on $u$;
            // computed using $l$ draws
        add $\langle x, \phi_x \rangle$ to RankedExamplePool;
    else
        remove $\{x, y\}$ from $\mathcal{U}'$;
    end
end
Acquire labels for top examples sorted by $\phi_x$
  descending as $\{\hat{y}_1, ..., \hat{y}_{\hat{n}}\}$
Set $\mathcal{U}'$ to be $(x_1, \hat{y}_1), ..., (x_{\hat{n}}, \hat{y}_{\hat{n}})$
Update model: draw the next training example from $\mathcal{U}'$
  over $\phi_x$
    get $\hat{x} = \text{argmax}_u\, 1 - P_\theta(\hat{y}|x)$; - # sample we are
  least confident of
    retrain learning model using $\{\hat{x}, \hat{y}\}$
return $\mathcal{U}'$
```

RESULTS

The task was binary intent classification in an active learning setting. The model was bootstrapped with 1,000 examples. The active learner then queries a data pool to get a batch of additional training examples to improve the model.



CONCLUSIONS

Our results show that by taking a small performance hit, we can achieve accuracy scores close to the baseline with a smaller annotation budget and stronger privacy guarantees.

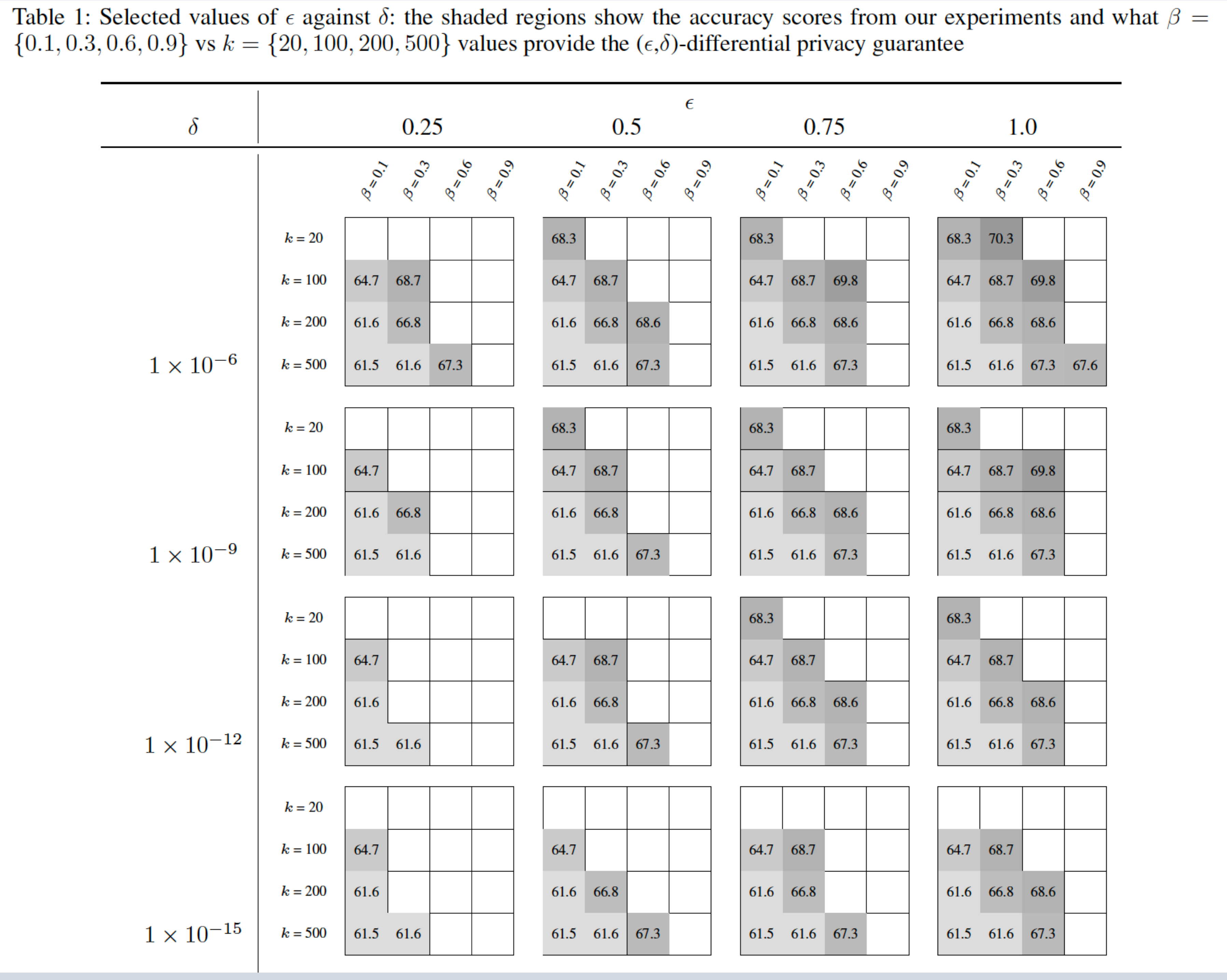


Table 1: Selected values of ϵ against δ: the shaded regions show the accuracy scores from our experiments and what $\beta = \{0.1, 0.3, 0.6, 0.9\}$ vs $k = \{20, 100, 200, 500\}$ values provide the (ϵ,δ)-differential privacy guarantee

δ		ϵ															
		0.25				0.5				0.75				1.0			
		β = 0.1	β = 0.3	β = 0.6	β = 0.9	β = 0.1	β = 0.3	β = 0.6	β = 0.9	β = 0.1	β = 0.3	β = 0.6	β = 0.9	β = 0.1	β = 0.3	β = 0.6	β = 0.9
1×10^{-6}	k = 20					68.3				68.3				68.3	70.3		
	k = 100	64.7	68.7			64.7	68.7			64.7	68.7	69.8		64.7	68.7	69.8	
	k = 200	61.6	66.8			61.6	66.8	68.6		61.6	66.8	68.6		61.6	66.8	68.6	
	k = 500	61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3	67.6
1×10^{-9}	k = 20					68.3				68.3				68.3			
	k = 100	64.7				64.7	68.7			64.7	68.7			64.7	68.7	69.8	
	k = 200	61.6	66.8			61.6	66.8			61.6	66.8	68.6		61.6	66.8	68.6	
	k = 500	61.5	61.6			61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3	
1×10^{-12}	k = 20									68.3				68.3			
	k = 100	64.7				64.7	68.7			64.7	68.7			64.7	68.7		
	k = 200	61.6				61.6	66.8			61.6	66.8	68.6		61.6	66.8	68.6	
	k = 500	61.5	61.6			61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3	
1×10^{-15}	k = 20																
	k = 100	64.7				64.7				64.7	68.7			64.7	68.7		
	k = 200	61.6				61.6	66.8			61.6	66.8			61.6	66.8	68.6	
	k = 500	61.5	61.6			61.5	61.6	67.3		61.5	61.6	67.3		61.5	61.6	67.3	

REFERENCES

[1] Burr Settles. Active Learning literature survey. 2010. Computer Sciences Technical Report

[2] Dwork, Cynthia, et al. "Calibrating noise to sensitivity in private data analysis. Theory of cryptography Conf. 2006.

[3] Li et al. On sampling, anonymization, and differential privacy or k-anonymization meets differential privacy.