Empirical privacy and empirical utility of anonymized data

Graham Cormode [#], Cecilia M. Procopiuc [#], Entong Shen [*], Divesh Srivastava [#], Ting Yu [*]

[#] *AT&T Labs – Research,*
{graham,magda,divesh}@research.att.com

[*] *North Carolina State University,*
{eshen,tyu}@ncsu.edu

***Abstract*— Procedures to anonymize data sets are vital for companies, government agencies and other bodies to meet their obligations to share data without compromising the privacy of the individuals contributing to it. Despite much work on this topic, the area has not yet reached stability. Early models (k-anonymity and ℓ-diversity) are now thought to offer insufficient privacy. Noise-based methods like differential privacy are seen as providing stronger privacy, but less utility. However, across all methods sensitive information of some individuals can often be inferred with relatively high accuracy.**

In this paper, we reverse the idea of a 'privacy attack,' by incorporating it into a measure of privacy. Hence, we advocate the notion of *empirical privacy*, based on the posterior beliefs of an adversary, and their ability to draw inferences about sensitive values in the data. This is not a new model, but rather a unifying view: it allows us to study several well-known privacy models which are not directly comparable otherwise. We also consider an empirical approach to measuring utility, based on a workload of queries. Consequently, we are able to place different privacy models including differential privacy and early syntactic models on the same scale, and compare their privacy/utility tradeoff. We learn that, in practice, the difference between differential privacy and various syntactic models is less dramatic than previously thought, but there are still clear domination relations between them.

I. Introduction

Increasingly, organizations must make available versions of the data they are collecting, whether for legal or business reasons. At the same time, data owners are under strong obligations not to compromise the privacy of the individuals represented in their data. As a result, there is a need to provide "anonymized" versions of data which balance these two requirements. Initial efforts for developing privacy models [1], [2], [3] focused on weakening (or breaking) the connection between "quasi-identifiers" and "sensitive values". These offer appealing and seemingly intuitive properties of the likelihood of certain facts holding in the original data. However, some connections can be reconstructed from the published data, using statistical inference and/or knowledge of the anonymization procedure [4], [5].

The differential privacy model [6], which has gained considerable support in the database community, imposes a conceptually different condition: its output is nearly identical (in a probabilistic sense), whether or not an individual contributes her data to the set. But this model is not immune to the same limitations of the prior models: an attacker can often draw accurate inferences about a person's sensitive information. This arises particularly when there are known correlations in the data [7], or when the published dataset exposes at least one previously unknown correlation between a sensitive value and the values of some other attributes [8].

Despite such challenges, the pressures to provide data are sufficiently strong that releases will occur, and hence we need to enhance our understanding of the tools that we currently have, imperfect as they may be. Despite the weaknesses discussed above, the resulting privacy breaches are not absolute: it is not the case that an adversary learns every private piece of information with absolute certainty. Instead, in these examples an adversary ends up with a set of (probabilistic) beliefs about individuals. In some cases, these may be damaging (the adversary may gain a strong belief about some individual), while in other cases it may be less so. Hence, our aim in this paper is to focus on *quantifying* the privacy impact of a data release. That is, we introduce the idea of incorporating a measure over "privacy breaches" into a definition of empirical privacy, and evaluating the corresponding empirical utility of the released data. We next elaborate on each of these notions.

Empirical privacy. In the race to propose ever stronger models, privacy guarantees have become increasingly obscure for both the data owner, and the average individual whose data they are trying to protect. This is a significant drawback in many cases: people are unlikely to contribute their data to surveys, or entrust organizations with their information, if the data aggregators cannot explain their privacy guarantees clearly. Data owners are hesitant to release data via complex algorithms if they cannot parse the guarantees that the algorithms provide, or compare the results from different methods. Given the above discussion, and the need for a simpler privacy explanation, we advocate the notion of *empirical privacy* as a measurement tool. Roughly speaking, it represents the precision with which the sensitive values of individuals can be inferred from released data. This is inspired by the fact that current models of privacy attack focus on a sophisticated adversary who can use the tools of statistical inference on released data [5], [8]. This is therefore not a new specific privacy model, but rather a measure of privacy and a unifying view: it allows us to study several well-known privacy models which are not directly comparable otherwise.

Empirical utility. In most cases, the utility of the released data is not part of the privacy model. Early work used "information loss metrics", which measure how much the original data was coarsened, but it is unclear how they correlate with any use of data [9]. The difficulty is that the notion of utility is ill-defined, depending on some unknown future use of the data. In this paper, we define the *empirical utility* to be the (relative) error of COUNT(*) queries with range conditions on the attributes, since such queries can essentially be used to describe the distribution of data and serve as the building blocks of more complex data analysis. They have been used in several experimental evaluations of prior work in isolation.

In this work, we bring together several well-known privacy models: k-anonymity, ℓ-diversity, t-closeness and ε-differential privacy. Although the models are quite different, and their theoretical guarantees cannot be directly compared, our notions of empirical privacy and empirical utility apply to all of them. Our notions of privacy and utility are distinct, dealing with single individuals and groups, respectively. That is, 'privacy' relates to the ability to learn about individuals, while 'utility' relates to the ability to learn aggregate statistics about large groups of individuals (this distinction is also made in [10]).

Our contributions. We propose a unifying framework for comparing privacy models, by formalizing the notions of empirical privacy and empirical utility. We provide an experimental study to compare several privacy models. Surprisingly, this shows that the difference between the various models is less dramatic than previously thought.

Background on privacy models. Privacy is a fast evolving research area, and a full survey of the field is beyond the scope of this paper (instead, see [11]). The class of deterministic models ensure that certain properties hold (deterministically) over the output. These privacy models make a distinction between the type of values present in a tuple, grouping them into quasi-identifiers and sensitive values. The assumption is that an individual's quasi-identifier values are either widely known, or can be easily determined; whereas the sensitive values are private and must be protected by the model. The general approach is to split tuples into groups and to redact their quasi-identifier values (e.g., via generalization or suppression), so that tuples in a group become indistinguishable when projected over the quasi-identifiers. We assume familiarity with well-known models, such as k-anonymity [1] (each group must have at least k tuples); ℓ-diversity [3] (each group must contain at least ℓ distinct and "well-represented" sensitive values); and t-closeness [12] (the distribution of sensitive values in each group is not too different from their global distribution). Meanwhile, randomized methods draw noise from appropriate distributions to perturb the output. In differential privacy, the probability of any property holding on the output must be approximately the same, whether or not an individual is present in the source data [6].

II. Empirical Privacy and Utility

We introduce a measure of privacy independent of any specific privacy model, and more generally interpretable by users. This measure is inspired by a widely adopted notion of a privacy breach: the correct posterior inferences of an adversary about sensitive values in the data. Let $\mathcal{A}$ be an anonymization mechanism, and $\mathcal{A}(\mathcal{D})$ be its output on a dataset $\mathcal{D}$. Let S denote the sensitive attribute of $\mathcal{D}$. To evaluate the empirical privacy we will use $\mathcal{A}(\mathcal{D})$ to build a model of the data so that, for each tuple $\tau \in \mathcal{D}$, we can compute a set of predictions for the sensitive values $\tau.S$, with an associated belief; i.e., we compute the pairs $(v_1, p_1), (v_2, p_2), \ldots$, where $0 < p_i < 1$ is the belief that $\tau.S = v_i$; and $\sum p_i \leq 1$. From these, we choose the value v_i with highest belief p_i as the prediction for $\tau.S$; the prediction confidence is p_i.

Definition 1: Let $\mathcal{D}$ be a dataset with sensitive attribute S, and $\mathcal{A}$ be an anonymization mechanism. The *average empirical privacy*, $\pi(\mathcal{A}(\mathcal{D}))$ is the fraction of tuples $\tau \in \mathcal{D}$ for which we do not predict the correct value $\tau.S$.

We note that variations of this definition are possible, e.g. restricting to the top-k tuples based on the confidence in their prediction. In our experiments, we report the empirical *privacy breach increase* $\beta(\mathcal{A}(\mathcal{D}))$, measured as

$$\beta(\mathcal{A}(\mathcal{D})) = (1 - \pi(\mathcal{A}(\mathcal{D})))/\rho - 1,$$

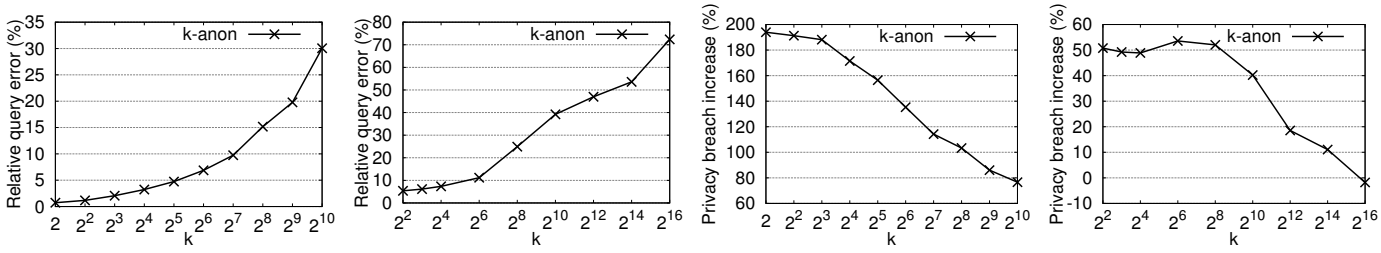
where $\pi(\mathcal{A}(\mathcal{D}))$ is the empirical privacy value given by Definition 1, and ρ is the accuracy of the baseline approach in which the most frequent sensitive value is always predicted. This is consistent with prior work that has used similar measures [16].

For utility, we measure the query accuracy for a given query workload.

Definition 2: Let $\mathcal{A}(\mathcal{D})$ be the output of an anonymization procedure on a dataset $\mathcal{D}$. Let $\mathcal{Q}$ be a query workload of COUNT(*) queries with range conditions on all attributes of $\mathcal{D}$. The relative error of a query $q \in \mathcal{Q}$ is the ratio $\mathrm{rel}(q) = \frac{|q(\mathcal{A}(\mathcal{D})) - q(\mathcal{D})|}{q(\mathcal{D})}$, where the notation $q(X)$ denotes the answer to q computed over set X. In the experiment section, we report the median relative error $\alpha(\mathcal{A}(\mathcal{D})) = \mathrm{median}(\{\mathrm{rel}(q) | q \in \mathcal{Q}\})$, given a query workload $\mathcal{Q}$. The *empirical utility* of $\mathcal{A}(\mathcal{D})$ (with respect to $\mathcal{Q}$) can be defined as the reciprocal of median relative error, i.e., $1/\alpha(\mathcal{A}(\mathcal{D}))$, so that larger values indicate higher utility of $\mathcal{A}(\mathcal{D})$.

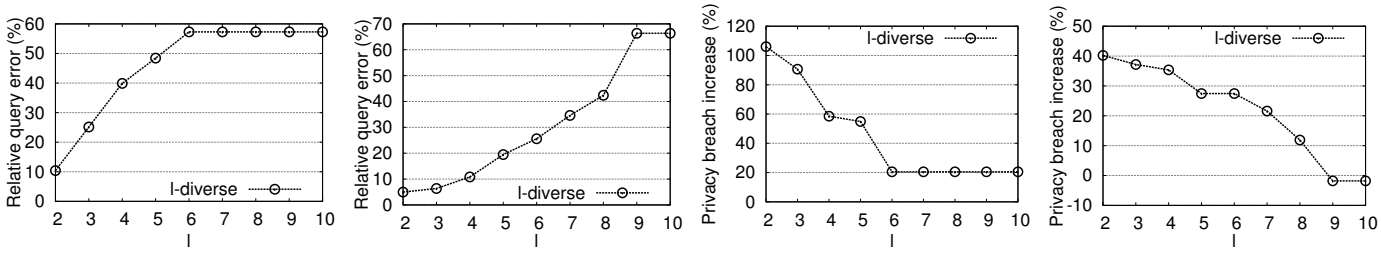
Anonymization techniques. In this paper, we focus on anonymization schemes that provide their output in the form of *spatial decompositions*. That is, they partition the space of attributes into different regions, and provide a description of the density of values within each region. This captures many prior works, e.g., the Mondrian approach [9] for k-anonymity, and the ε-differentially private kd-trees and quad-trees studied in [13], [14], [15]. We omit detailed description of how these are built. Often the regions in the spatial decomposition form a hierarchical division of the space. For each leaf in the spatial decomposition tree, a histogram describes the sensitive values associated with the points in the leaf region. For the deterministic models, this contains the exact counts; for probabilistic models, noise is added to the histogram counts.

Attack model. We now describe how we instantiate our measure of empirical privacy, based on the beliefs of an adversary



(a) Adult-query (b) Census-query (c) Adult-predict (d) Census-predict

Fig. 1. Impact of k on k-anonymity



(a) Adult-query (b) Census-query (c) Adult-predict (d) Census-predict

Fig. 2. Impact of ℓ on ℓ-diversity

who sees the output of an anonymization mechanism. However, there are many possible adversaries, depending on their prior beliefs and model of inference. We focus on an adversary who uses the output $\mathcal{A}(\mathcal{D})$ to build a classifier to attack anonymized data. This notion is at the heart of some prior works [16], [5], [8]. It is appealing: it does not rely on making explicit the prior beliefs of the adversary (requiring some external knowledge or domain-specific reasoning). Rather, it allows us to provide a general mechanism to quantify empirical privacy. There are still many possible choices of classifier to instantiate. However, our empirical experience suggests that different classifiers achieve similar levels of empirical privacy, varying only by low single digit percentages.

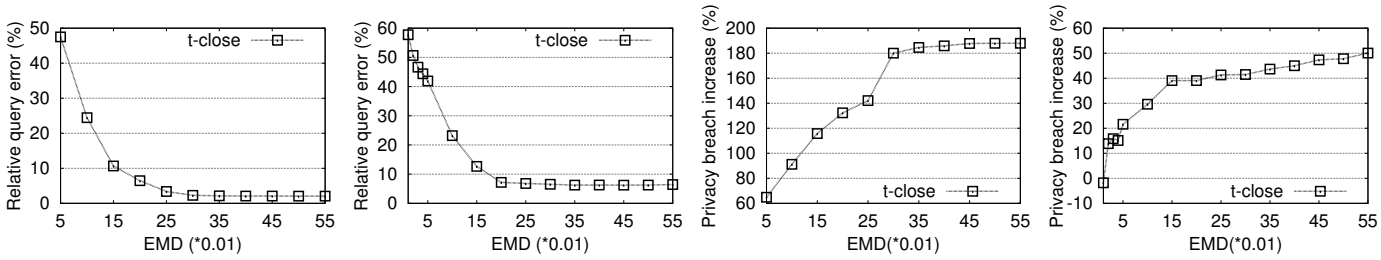
Given anonymized data in the spatial decomposition format, we now determine how to instantiate a classifier to compute the empirical privacy of the data. This is somewhat non-trivial, since this is not the usual classification problem: typically, one expects to see training data in the form of specific examples and labels. In our setting, we have regions instead of examples, and a multiset of labels associated with each region. We describe how to build a naive Bayes classifier, which is chosen by prior work to apply to anonymized data [16], [5], [8]. It has been observed to achieve good accuracy over many data instances. The classifier aims to find correlations between each quasi-identifier and the sensitive attribute. The parameters of the classifier are the conditional probabilities, $\mathsf{Pr}[t.j = u | t.S = v]$: the probability that the jth attribute has value u, given that the sensitive value is v. We also need the global distribution of sensitive attributes, $\mathsf{Pr}[t.S = v]$. Then, given an individual τ, the classifier provides the beliefs p_i as:

$$p_i = \frac{\mathsf{Pr}[t.S = v_i] \prod_{j=1}^{d} \mathsf{Pr}[t.j = \tau.j | t.S = v_i]}{\sum_v \mathsf{Pr}[t.S = v] \prod_{j=1}^{d} \mathsf{Pr}[t.j = \tau.j | t.S = v]}.$$

To derive the necessary conditional distributions when the anonymized data is in the output format of a spatial decomposition we apply a simple kernel approach: given a leaf containing a histogram of sensitive values, we use a uniformity assumption and treat each data point as spread uniformly along the extent of the leaf in each quasi-identifier dimension. Specifically, suppose that in some leaf we have n occurrences of sensitive value v, and that the leaf covers the range $[x, y]$ along attribute j. Then we treat this as a collection of (weighted) tuples t having $t.S = v$, where the tuple with value $t.j \in [x, y]$ has weight $n/|[x, y]|$. Summing all these kernels over all leaves gives us (after rescaling) the joint probability distribution $\mathsf{Pr}[t.S = v, t.j = u]$. From this, we derive the conditional distribution $\mathsf{Pr}[t.j = \tau.j | t.S = v]$ from the identity

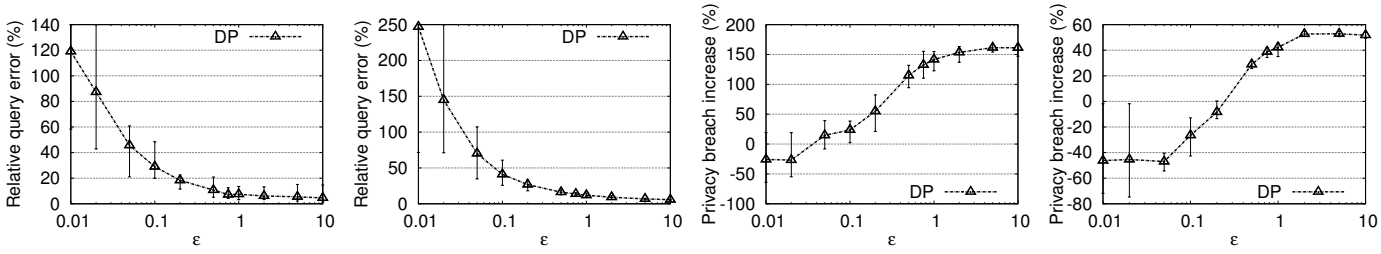
$$\mathsf{Pr}[t.j = \tau.j | t.S = v] = \frac{\mathsf{Pr}[t.S = v, t.j = \tau.j]}{\mathsf{Pr}[t.S = v]}$$

For data produced via differential privacy, we note that it is possible to achieve negative histogram counts associated with some values of $t.S$, due to the random noise distribution. This is remedied by rounding such counts up to zero (the most likely true value). We remark that our approach can be extended by, e.g., adopting different kernels for the smoothing; combining multiple attributes to build more sophisticated Bayes classifiers; or other standard variations on building a classifier. However, this baseline method is generally applicable, and already gives a sufficiently accurate classifier.



(a) Adult-query (b) Census-query (c) Adult-predict (d) Census-predict

Fig. 3. Impact of t on t-closeness



(a) Adult-query (b) Census-query (c) Adult-predict (d) Census-predict

Fig. 4. Impact of privacy budget ϵ on differential privacy

III. Experimental Study

We begin by exploring design parameters and their impact on empirical privacy and utility for each model: the maximum number of individuals in a group, k, for k-anonymity, and the parameters ℓ for ℓ-diversity, t for t-closeness and ϵ for differential privacy. We then compare all these models together by showing how utility changes with respect to various privacy levels. We also investigate the impact of other design choices, such as the maximum tree height and the size of the dataset.

Experiment Setup. Experiments were performed on two real data sets containing demographic data: the *Adult* dataset from the UCI Machine Learning repository[1] with 30,162 tuples; and the 2009 *Census* microdata we extracted from IPUMS USA [17] with 100,000 tuples. For Adult, occupation is the SA, and workclass, education, sex, work hours and income level are used as QIs. For Census, age, insurance state, education and occupation are the QIs, and salary-class is the SA.

We draw a workload of 2,000 queries with non-zero true answers, each of which covers half of the domain of each attribute. The median selectivity is about 13% for *Adult* data and 8% for *Census* data. For a workload of queries, the median relative error $\alpha(\mathcal{A}(\mathcal{D}))$ was reported as defined in Section II.

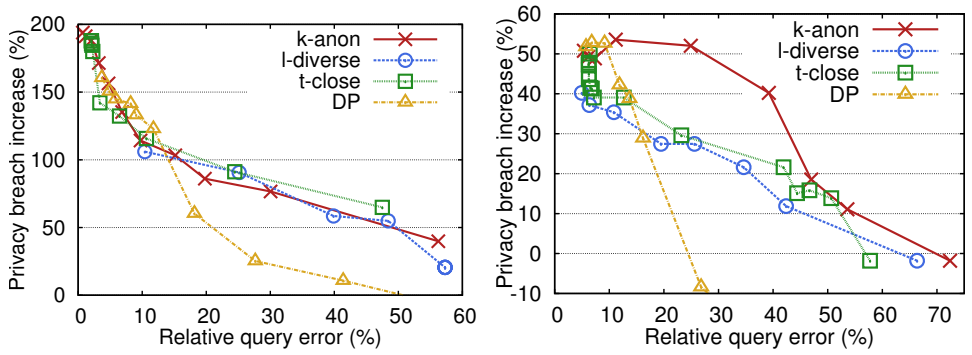
The relative increase in prediction accuracy $\beta(\mathcal{A}(\mathcal{D}))$ was reported, as described in Section II. For both datasets, the baseline accuracy is 11%. Occasionally, the accuracy of the classifier is worse than this baseline, leading to negative values of privacy breach increase in some plots. All experiments were conducted on a 3.00GHz CPU with 4GB RAM, so the data fits easily in memory. We implemented our anonymization trees in Python 2.6 with scientific package Numpy.

[1] http://archive.ics.uci.edu/ml/datasets/Adult

Privacy parameters. We first investigate the impact of the privacy parameter of each anonymization model independently. Figure 1 and Figure 2 show the impact of k on k-anonymity and ℓ on ℓ-diversity respectively. Since no noise is introduced for the deterministic models, the error in query answering comes solely from the uncertainty introduced by the rectangles of the leaves of the spatial decomposition tree. To answer queries, we make a standard uniformity assumption, and estimate the fraction of points covered in each leaf via the fraction of the rectangle covered by the query. As expected, the query accuracy suffers significantly with the increase of k and ℓ. In particular, when $\ell \geq 6$ for *Adult* dataset ($\ell \geq 8$ for *Census* dataset), it is not possible to achieve ℓ-diversity on the data, hence the tree does not split beyond the root level. Comparing these two models, ℓ-diversity is harder to achieve than k-anonymity, and its query and prediction accuracy decrease sharply with each increment of ℓ.

For empirical privacy, the effectiveness of the Naive Bayes classifier can be seen by the more than 180% (respectively, 50%) increase in prediction accuracy for k-anonymity on the *Adult* (respectively, *Census*) data when $k \leq 8$. Even for a large k, e.g. $k = 1024$, the classifier performs 50% better than the baseline case for both datasets. It is noticeable in Figure 2(c) and 2(d) that the lower bound of increase in prediction accuracy is about 20% for *Adult* data and about zero for *Census* data. This is the case when the classifier is applied directly on the root node.

Besides k-anonymity and ℓ-diversity, we also tested a variant of t-closeness built on top of k-anonymity ($k = 8$). It ensures that the histogram for any node remains close



(a) Adult dataset (b) Census dataset

Fig. 5. Empirical privacy breach vs empirical utility

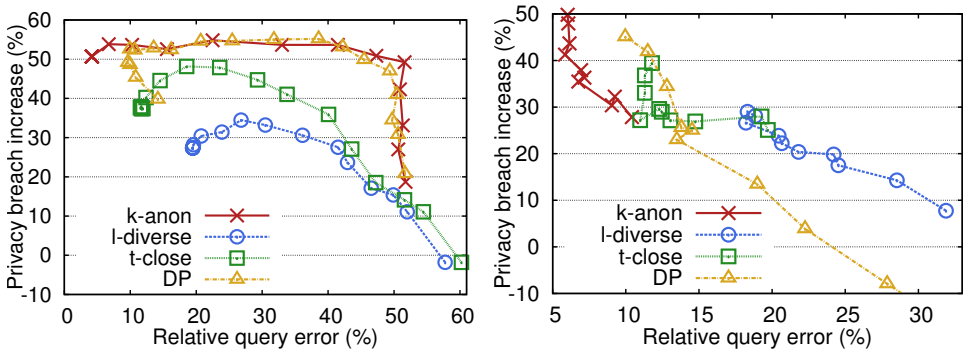
to the histogram of the sensitive attribute in the original data set. Here we use the Earth Mover's Distance (EMD) to measure the distance between two histograms. When building the tree, nodes having EMD (compared to root) larger than the threshold t will not be split further. Figure 3 shows the impact of this threshold. Clearly, a smaller EMD threshold will incur larger leaf nodes and affect the accuracy of both query answering and prediction. The impact of this closeness requirement levels off when the EMD threshold is larger than 0.35.

For differentially private algorithms, the value of the privacy budget ϵ plays a key role by setting the upper bound of the ratio of one's inference when an individual tuple is present in the data set or not. A smaller ϵ results in higher noise and thus more privacy. Prior work typically sets $\epsilon \leq 1$, but in some cases has tolerated larger values of ϵ (e.g., [18]). Here we vary ϵ from 0.01 to 10 in order to explore a wide spectrum of values. The experiments on differential privacy exhibit high variance due to the randomness involved (especially for small ϵ). Therefore we provide results averaged over 8 runs with distinct seeds of randomness. As shown in Figure 4, ϵ is inversely related to the privacy guarantees of differential privacy. Using a very small ϵ (e.g., 0.01), the noise added may even outweigh the underlying signal, providing the strongest privacy seen as approximately 100% of relative errors in query answering and worse-than-baseline predictions for both data sets. At the other end of the privacy-utility tradeoff, an extremely large privacy budget (e.g., $\epsilon = 10$) adds almost zero noise (in variance) to the original data, providing little query error at the cost of much compromise in privacy.

Comparison across anonymization models. The unifying framework introduced in Section II enables us to do a head-to-head comparison of deterministic and differentially private anonymizations. Inspired by the ROC curve frequently used in machine learning, we combine privacy and utility into a single plot as shown in Figure 5: the x-axis and y-axis represent empirical utility and empirical privacy breach respectively. This enables us to visualize the privacy-utility tradeoff of various models in one plot. We generate different (utility, privacy) pairs by varying privacy parameters for all the models, i.e., the k, l, t, ε values. For example, the data points on the 'DP' curve correspond to the relative query errors and prediction accuracy increases when varying the parameter ε from 0.01 to 10.

Notice that the bottom-left corner of the utility-privacy graph is where an ideal anonymization curve would be—high utility coupled with high privacy. Due to the privacy-utility tradeoff, however, most anonymization models will follow a more diagonal path from top-left (high utility, low privacy) to bottom-right (low utility, high privacy). As seen in both Figure 5(a) and 5(b), all anonymization models converge at the top-left corner (except ℓ-diversity, for which ℓ is at least 2), where privacy is sacrificed for high utility. This indicates that if low relative query error is needed, a large privacy compromise seems inevitable regardless of the anonymization model one may choose. In this range, all methods produce a quite similar local density model with little or no noise on counts. Interestingly, they start to diverge when moving towards the other end of the main diagonal. Since the bottom-left corner is the ideal region to fall into, the performance of an anonymization model can be judged by the relative distance to the origin of coordinates.

In Figure 5(a), t-closeness may be the preferred model when high utility (relative error $< 10\%$) is required, providing the smaller increase in prediction accuracy. If privacy guarantee is of more concern, e.g., requiring less than 100% accuracy increase, the advantage of differential privacy definitely shows up and dominates. A similar trend can be observed on the *Census* dataset in Figure 5(b). In particular, k-anonymity performs poorly on this dataset: relative error increases fairly consistently from 5% to 40%, but this only affects the classifier accuracy by about 10%. Only when the relative error grows very high does the classifier accuracy drop off. The ℓ-diversity and t-closeness models, which both restrict the distribution in each region, are reasonably similar. Again, the curve of differential privacy has a higher slope, which helps to reduce the accuracy of classifier prediction. If one can tolerate 25% relative error, differential privacy can provide nearly zero classifier accuracy increase over the baseline, while other deterministic models give poor utility (about 60% in relative



(a) privacy-utility as height varies

(b) privacy-utility as size varies

Fig. 6. Varying model height and data size

error) to achieve the same level of privacy protection.

We make several observations. First, no anonymization model is able to win hands down across the whole range. Specifically, despite being widely regarded as a 'superior' model, differential privacy provides the same or even worse empirical privacy protection when single-digit relative error (high utility) is required. In this region either ℓ-diversity or t-closeness can outperform differential privacy under these metrics as shown in both plots. However, if a strong privacy guarantee is required, differential privacy does not lose too much query accuracy. Meanwhile, using $\varepsilon = 1$ (shown as the fourth triangular marker from the left) may not be advisable: the empirical privacy here is not very high, and there are syntactic approaches providing better privacy-utility tradeoff in this region. At least for these two datasets, $\varepsilon \leq 0.5$ seems to be a suitable setting for differential privacy.

Impact of other parameters on the trade-off. Figure 6(a) shows the utility-privacy tradeoff curves, over different settings of the maximum height of the spatial decomposition tree, on the Census dataset. It shows that for these particular parameter settings, k-anonymity (with $k = 8$) and differential privacy (with $\varepsilon = 1$) provide almost the same privacy-utility tradeoff for a wide range of h values tested.

Figure 6(b) shows the effect on the curve of varying the size of the data 10% to 100% of the original input. Increasing the dataset size without changing other parameters increases the accuracy of both query answering and the classifier. We observe an interesting dependence on the size of data available: when less data is provided, differential privacy occupies the bottom right of the plot (more privacy, less utility); but when more data is provided, it is dominated by k-anonymity, in the region of less privacy but more utility.

IV. Concluding Remarks

By taking a pragmatic approach to anonymization, we are able to take a holistic view of the variety of different privacy models that have been proposed. The bottom line is that differential privacy often provides the best empirical privacy for a fixed (empirical) utility level, but for more accurate answers it can be preferable to adopt a method like t-closeness or ℓ-diversity (with correspondingly higher privacy risk). This matches our intuitive expectation, but quantifies it more rigorously. Further, we see that by these measures, the difference between the methods is not so large. This suggests different use-cases: when releasing data to a third party (say, an external data analysis company), differential privacy is the current method of choice. But when releasing data to a more trusted entity (say, a different department within the same organization), ℓ-diversity suffices to prevent trivial data leakages while preserving more of the utility.

References

[1] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *ICDE*, 2005.

[2] X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preservation," in *VLDB*, 2006.

[3] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," in *ICDE*, 2006.

[4] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, "Minimality attack in privacy preserving data publishing," in *VLDB*, 2007, pp. 543–554.

[5] D. Kifer, "Attacks on privacy and deFinetti's theorem," in *SIGMOD*, 2009.

[6] C. Dwork, "Differential privacy," in *ICALP*, 2006, pp. 1–12.

[7] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *SIGMOD*, 2011.

[8] G. Cormode, "Individual privacy vs population privacy: Learning to attack anonymization," in *KDD*, 2011.

[9] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *ICDE*, 2006.

[10] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *KDD*, 2009, pp. 517–526.

[11] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, *Privacy-Preserving Data Publishing*, ser. Foundations and Trends in Databases. NOW publishers, 2009.

[12] L. N. Li, L. T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," in *ICDE*, 2007.

[13] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, "Private record matching using differential privacy," in *EDBT*, 2010.

[14] Y. Xiao, L. Xiong, and C. Yuan, "Differentially private data release through multidimensional partitioning," in *SDM Workshop at VLDB*, 2010.

[15] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, "Differentially private spatial decompositions," in *ICDE*, 2012.

[16] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *KDD*, 2008.

[17] S. Ruggles, J. Alexander, K. Genadek, R. Goeken, M. Schroeder, and M. Sobek, "Integrated public use microdata series: Version 5.0." *Minneapolis, MN: Minnesota Population Center, 2010*.

[18] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *ICDE*, 2008.