Research Statement

Jun Zhang

1. INTRODUCTION

My PhD research focuses on **data privacy**. This research is motivated by practical need of protecting individual privacy in the big data era, where private information is collected, analyzed, and disseminated in a massive scale. Specifically, organizations nowadays often allow third parties to perform business analysis or provide services using personal data. For instance, social network applications are often allowed to use individual social data to provide services, such as gaming, car-pooling, online deals. As another example, retailers often share data about their customers' purchasing behaviors with product merchants for more effective and targeted advertising. While sharing information can be highly valuable, companies that provide accesses to private data have to be extremely careful in controlling the risk of disclosure. Yet, recent years have witnessed quite a few incidents in which organizations fail to protect privacy in the release of sensitive data, e.g., the AOL search log scandal in 2006 and the de-anonymization of Netflix prize data in 2007.

2. PROBLEM DESCRIPTION

Releasing sensitive information while preserving individuals' privacy has been an active research subject for decades. The recently proposed notion of ε-*differential privacy* is rapidly emerging as the state-of-the-art scheme for this purpose, due to its strong privacy guarantees and robustness against adversaries with background knowledge. Given a sensitive dataset, ε-differential privacy requires that any information published from the data must be *randomly perturbed* to prevent inference of sensitive information. In particular, it ensures that, even if a potential adversary knows the exact details of all but one record in the data, it would still be difficult for the adversary to infer the information about the remaining record. The objective of my PhD research is to develop new techniques for publishing data under differential privacy, to (i) tackle previously unsupported publication tasks, (ii) provide improved data utility over the state of the art, and (iii) devise general mechanisms that will benefit the design of solutions for a board range of privacy problems.

3. PROGRESS

So far I have published four papers on my PhD research topic, and the first two of them focus on solving model fitting and optimization problems under differential privacy. The first paper is titled "Functional Mechanism: Regression Analysis under Differential Privacy", and it was published in VLDB 2012. The paper presents an *objective perturbation* technique for differentially private linear and logistic regressions. The main idea is to enforce ε-differential privacy by perturbing the objective function of the optimization problem, and then solving the regression task under the perturbed function. The perturbation of the objective function is highly non-trivial as it requires analyzing the sensitivity of the function's output with respect to changes in an arbitrary record in the input. We show that this issue can be alleviated by adopting a truncated Taylor series to approximate the objective function. The resulting solution is shown to outperform alternative solutions in terms of data utility.

My second paper is titled "PrivGene: Differentially Private Model Fitting Using Genetic Algorithms", and it was accepted by SIGMOD 2013. The paper presents *PrivGene*, which is a solution that incorporates the classic genetic algorithm with differential privacy to solve model fitting problems. PrivGene starts with a set of random parameter vectors, and iteratively improves them by applying minor perturbation to the elements of the vectors (this mimics mutations in natural evolutions) and then filtering out the vectors that do not fit the given data well (which mimics the selection process in natural evolutions). The filtering of vectors is performed with *enhanced exponential mechanism (EEM)*, which is a novel technique that improves over the standard *exponential mechanism* by exploiting special properties of the selection operations.

The third work is "PrivBayes: Private Data Release via Bayesian Networks", accepted by SIGMOD 2014. *PrivBayes* is a differentially private method for releasing high-dimensional data. PrivBayes first constructs a Bayesian network which (i) provides a succinct model of the correlations among the attributes in the input data and (ii) allows to approximate the distribution of data using a set of low-dimensional marginals. After that, PrivBayes injects Laplace noise into each marginal to ensure differential privacy, and then uses the noisy marginals and the Bayesian network to construct an approximation of the tuple distribution in the input data. Finally, PrivBayes releases a synthetic data, whose tuples are sampled from the approximate distribution. Private construction of Bayesian networks turns out to be significantly challenging, and we introduce a novel approach that uses a surrogate function for mutual information to build the model more accurately.

The most recent work "Private Release of Graph Statistics using Ladder Functions" is published in SIGMOD 2015. In this paper, we introduce a new technique for producing differentially private output, which is applied to the problem of producing subgraph counts. The technique builds upon ideas from differential privacy, specifically notions of the "sensitivity" of a function, and sampling possible output values using the exponential mechanism and a carefully tuned quality function. We combine these in a new way to define a class of "ladder functions", which provide the optimal sampling distribution for outputs. When applied to subgraph counting queries, the results are efficient to compute, and improve substantially over prior work on these problems, where applicable. In addition, the results of our algorithms can be used to estimate the parameters of suitable graph models, allowing synthetic graphs to be sampled.