Interactive Anonymization of Sensitive Data

Xiaokui Xiao, Guozhang Wang, Johannes Gehrke
Department of Computer Science
Cornell University
Ithaca, New York
{xiaokui, guoz, johannes}@cs.cornell.edu

ABSTRACT

There has been much recent work on algorithms for limiting disclosure in data publishing. However, these algorithms have not been put to use in any comprehensive, usable toolkit for practicioners. We will demonstrate CAT, the Cornell Anonymization Toolkit, designed for interactive anonymization. CAT has an interface that is easy to use; it guides users through the process of preparing a dataset for publication while limiting disclosure through the identification of records that have high risk under various attacker models.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration—*Security, integrity, and protection*

General Terms

Design, Security

Keywords

Data anonymization, l-diversity

1. INTRODUCTION

Organizations (such as the Census Bureau or hospitals) collect large amounts of personal information. This data has high value for the public, for example, to study social trends or to find cures for diseases. However, careless publication of such data poses a danger to the privacy of the individuals who contributed data.

There has been much research over the last decades on methods for limiting disclosure in data publishing; in particular, the computer science community has made important contributions over the last ten years. The research in this area investigated various adversary models and proposed different anonymization techniques that provide rigorous guarantees against attacks. However, to the best of our knowledge, none of these techniques has so far been implemented as part of a usable tool[1]. This is mainly due to the non-interactive nature of these techniques: The only interface they provide to data publishers is a set of parameters that controls the degree of privacy protection to be enforced in the anonymized data. The publishers, however, seldom have enough knowledge to decide appropriate values for the parameters; setting these values requires not only a deep understanding of the underlying privacy model but also a thorough understanding of possible adversaries. Furthermore, even if the data publisher had such knowledge, she much prefers a interactive anonymization process instead of fixing the algorithm and its parameters before seeing the anonymized output data. The data publisher will select the final anonymized version of the data only after she has explored the space of anonymization parameters and adversary models. Existing anonymization techniques have not been put into such a progressive, user-centric anonymization process.

In this demonstration, we bring the theory of data anonymization to practice. We developed CAT, the Cornell Anonymization Toolkit, that not only incorporates the state-of-the-art formal privacy protection methods, but also provides an intuitive interface that can interactively guide users through the data publishing workflow. CAT was designed with two objectives in mind. First, the toolkit should help users to acquire an intuitive understanding of the disclosure risk in the anonymized data, so that they can make educated decisions on releasing appropriate data. Second, the toolkit should offer the users full control of the anonymization process, allowing them to adjust various parameters and to examine the quality of the anonymized data (in terms of both privacy and utility) in a convenient manner. To the best of our knowledge, this is the first effort that employs existing anonymization techniques to provide a practical tool for data publication.

2. SYSTEM OVERVIEW

Figure 1 illustrates the major components of our system. The *anonymizer* uses an algorithm that, given some user-defined parameters, produces anonymized data that adheres to a user-selected privacy model. We anonymize data using *generalization* [4], which transforms attribute values of non-sensitive attributes (e.g., gender, date of birth, ZIP code) in the data into values ranges, so as to prevent an adversary from identifying individuals by linking these attributes with public available information. Currently, our system implements the *Incognito* algorithm [1] and the *l-diversity* model [2]. To ensure responsiveness, the dataset to

[1]One algorithm has been adopted by the CENSUS Bureau for data publication of OnTheMap Version 3.0 [3]. However, the anonymization process was performed by experts that ran different parameter values of the algorithm instead of a manipulation of the data with a software tool.

SIGMOD'09, June 29–July 2, 2009, Providence, Rhode Island, USA.
ACM 978-1-60558-551-2/09/06.

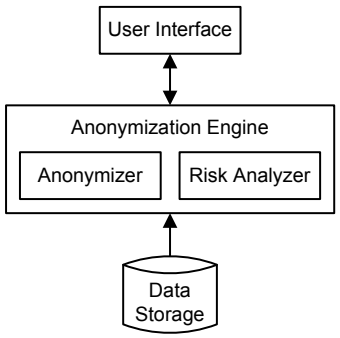

Figure 1: System architecture

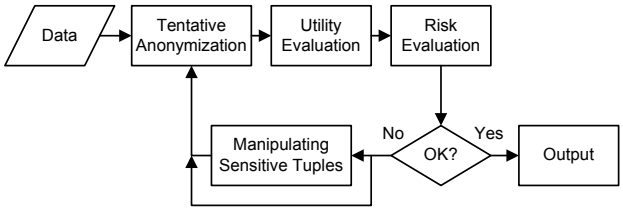

Figure 2: Anonymization process

be anonymized is kept in main memory, and all algorithms run against this main-memory resident data.

In addition to the anonymizer, we have a *risk analyzer* for evaluating the disclosure risks of records in anonymized data, based on user-specified assumptions about the adversary's background knowledge that can be specified through the user interface. Following the l-diversity model, we consider that the adversary may have information of the non-sensitive attributes of every individual, as well as several pieces of additional knowledge about the sensitive attributes. Each of these pieces of knowledge is modeled as a *negated atom*, i.e., a statement declaring that an individual is not associated with a certain sensitive attribute, such as "Alice does not have diabetes" or "Bob does not have cancer." We quantify the disclosure risk of an individual as the adversary's posterior probability of inferring the correct values of the sensitive attributes of the individual, after combing the anonymized data with the background knowledge.

3. DEMONSTRATION DESCRIPTION

We will demonstrate our system by showing the process of anonymizing a real dataset, as illustrated in Figure 2. We begin by loading the dataset into our system, upon which the tuples in the datasets will be shown in the upper-left panel of the user interface (see Figure 3). After that, we interact with the system to produce an anonymized table, by repeating the following four steps.

Step 1: Preliminary Anonymization. We first visit the middle-right panel, where there are two sliders that control the two parameters l and c of the l-diversity algorithm. Suppose that we do not have a clear idea of how these parameters should be set. Then, we simply select some initial values for l and c, and click the "Generalize" button. The system now computes a new anonymized table which serves as a starting point of our anonymization process. In the steps that follow, we will evaluate the quality of this anonymized table in terms of both privacy and utility. In case the table is unsatisfactory, we can refine it by adjusting the values of l and c.

	male	female	total
married	284421	48590	333011
divorced	37453	56581	94034
widowed	13546	61549	75095
single	48105	49755	97860
total	383525	216475	600000

Table 1: Contingency table

Step 2: Utility Evaluation. To get an understanding of the utility of an anonymization, we first click the "Contingency Tables" tab in the lower-left panel, to compare the *contingency tables* that correspond to the original and anonymized data, respectively. Specifically, a contingency table is a table that shows the frequencies for combinations of two attributes. For example, Table 1 illustrates a contingency table of gender and marital status. Intuitively, contingency tables show correlations between pairs of attributes. By examining the changes in the contingency tables before and after anonymization, we can get an idea of how the anonymization affects the characteristics of the data beyond looking at individual attributes. The two combo boxes in the top of the lower-left panel enable us to specify the two dimensions of the contingency tables.

After that, we will click on the "Density Graphs" tab, and the system will depict two density graphs that correspond to the contingency table, as shown in Figure 3. This provides us a more intuitive way to evaluate the differences between the original and anonymized data. In general, the more similar the graphs are, the more useful information is retained in the anonymized table.

Step 3: Risk Evaluation. We can now evaluate the privacy protection provided by the anonymized data. We begin by visiting the lower-right panel, and specify the amount of background knowledge that the adversary is expected to have. For example, in Figure 3, if we consider that the adversary is able to learn the ages of the individuals from public available information, then we can specify such knowledge of the adversary by putting a tick in the checkbox associated with "Background Knowledge" about "Age." In addition, we can use the slider in the bottom of the panel to define the number of negated atoms that the adversary may have about the sensitive attribute.

Once the background knowledge of the adversary is decided, we click the "Evaluate Risk" button, which will trigger an update in the upper-left panel. The system first calculates the disclosure risk of every record in the dataset based on the background knowledge. Thus makes the risks of the tuples available. For example, in Figure 3, the first tuple has a 4% disclosure risk, which means that, an adversary with the specified background knowledge would have 4% confidence to infer the income of the individual corresponding to the first tuple. In addition, the system also plots a histogram on the upper right panel that illustrates the distribution of the disclosure risks of all individuals in the dataset. For the

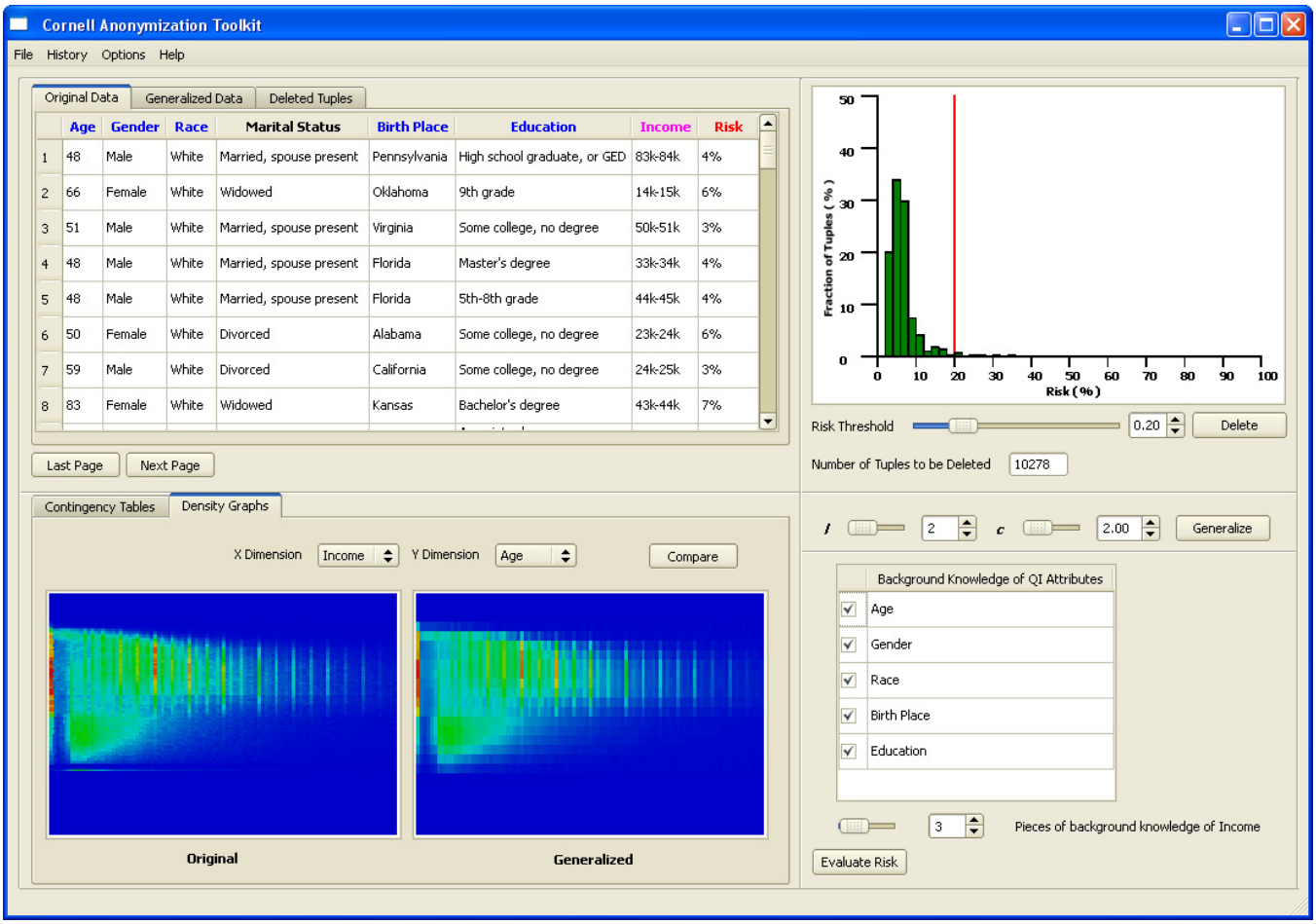

Figure 3: User interface

case in Figure 3, the histogram shows that the adversary has less than 20% confidence to infer the incomes of most individuals.

After inspecting the disclosure risks of the tuples, we have an intuitive understanding of the amount of privacy that is guaranteed by the anonymized table. In case both the privacy guarantee and the utility of the table are deemed sufficient, we request the system to output the table. Otherwise, we can move on to the next step.

Step 4: Manipulating Sensitive Tuples. In this step, we have the option of applying special treatment to special records in the table, i.e., records whose disclosure risks are much higher than most other tuples. Such tuples could be outliers in the dataset, and their existence may severely degrade the quality of anonymization in case we would not treat them separately. To eliminate such tuples, we first specify a threshold using the slider in the upper-right panel, and then click the "Delete" button to remove all those tuples whose disclosure risks are above the threshold. All deleted tuples can be reviewed in the "Deleted Tuples" tab of the panel, and can be restored whenever necessary.

We can now return to Step 1 and re-adjust the parameters in the middle-right panel to generate a new anonymized table. We apply this process iteratively until we obtain a satisfactory anonymization.

4. ACKNOWLEDGMENTS

This work was in part supported by NSF Grant CNS-0627585. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

5. REFERENCES

[1] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD*, pages 49–60, 2005.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *TKDD*, 1(1), 2007.

[3] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *ICDE*, pages 277–286, 2008.

[4] P. Samarati. Protecting respondents' identities in microdata release. *TKDE*, 13(6):1010–1027, 2001.