SECRETA: A System for Evaluating and Comparing RElational and Transaction Anonymization algorithms*

Giorgos Poulis
University of Peloponnese
poulis@uop.gr

Aris Gkoulalas-Divanis
IBM Research-Ireland
arisdiva@ie.ibm.com

Grigorios Loukides
Cardiff University
g.loukides@cs.cf.ac.uk

Spiros Skiadopoulos
University of Peloponnese
spiros@uop.gr

Christos Tryfonopoulos
University of Peloponnese
trifon@uop.gr

ABSTRACT

Publishing data about individuals, in a privacy-preserving way, has led to a large body of research. Meanwhile, algorithms for anonymizing datasets, with relational or transaction attributes, that preserve *data truthfulness*, have attracted significant interest from organizations. However, selecting the most appropriate algorithm is still far from trivial, and tools that assist data publishers in this task are needed. In response, we develop SECRETA, a system for analyzing the effectiveness and efficiency of anonymization algorithms. Our system allows data publishers to evaluate a specific algorithm, compare multiple algorithms, and combine algorithms for anonymizing datasets with both relational and transaction attributes. The analysis of the algorithm(s) is performed, in an interactive and progressive way, and results, including attribute statistics and various data utility indicators, are summarized and presented graphically.

1. INTRODUCTION

Publishing data about individuals is essential for applications, ranging from marketing to healthcare. Several marketing studies, for example, seek to find product combinations that appeal to customers with specific demographic profiles, while a large class of medical studies aims to discover associations between patient demographics and diseases. To enable these applications, data must be published in a way that preserves privacy and utility.

Towards this goal, numerous algorithms that prevent the disclosure of individuals' private and sensitive information, while maintaining *data truthfulness* (i.e., generate data that can be analyzed at a record level), have been proposed [4,6,7,10]. These algorithms work by transforming attribute values in a dataset (e.g., replacing them with more general values), and are applicable to either relational or transaction (set-valued) attributes. For example, an individual's year of birth is modeled as a relational attribute, while his/her purchased items are modeled as a transaction attribute. Furthermore, these algorithms can be combined, using a recent approach [9], to anonymize datasets with both relational and transaction attributes, referred to as RT-datasets.

While there is a growing interest for publishing protected and truthful data from governmental [8] and industrial organizations [1], selecting the most appropriate algorithm, for a given dataset and publishing scenario, remains a challenging and error-prone task. This is because both the effectiveness and efficiency of algorithms depend on: (a) *data characteristics* (e.g., the distribution of values in an attribute), (b) *various input parameters which affect the level of privacy and utility* (e.g., hierarchies that govern data transformation), and (c) *data utility requirements* (e.g., the need to accurately answer a certain query workload, or to adhere to constraints on the way values are transformed).

To assist data publishers in this task, we propose SECRETA, the first system for evaluating and comparing anonymization algorithms for relational, transaction, and RT datasets. Our system integrates 9 popular algorithms under a common, benchmark-oriented framework, and it allows data publishers to apply and analyze the performance of one or more of these algorithms. SECRETA operates in two modes, namely *Evaluation* and *Comparison*.

The Evaluation mode can be used to configure and evaluate the effectiveness of a given algorithm, with respect to data utility and privacy, as well as its efficiency. For capturing data utility, we employ several information loss measures [7, 12] and support data utility requirements. These requirements can be expressed using queries and/or *utility constraints* [7], which are specified by data publishers or generated automatically. Furthermore, SECRETA enables the use of 20 different combinations of algorithms to anonymize RT-datasets. The selection and management of these combinations is performed in an intuitive way that allows preserving different aspects of data utility.

The Comparison mode offers data publishers the ability to design and execute benchmarks for comparing multiple anonymization algorithms. These benchmarks facilitate an interactive and progressive comparison of sets of algorithms, with respect to their utility and efficiency. The results of the comparative analysis are summarized and presented graphically, allowing for fast and intuitive understanding of the effectiveness and efficiency of different algorithms.

To our knowledge, SECRETA is the only system that permits a comprehensive evaluation and comparison of recent anonymization techniques. The Cornell Anonymization Toolkit [11] demonstrates a single algorithm for relational data, also supported by SECRETA, while TIAMAT [3] does

*More details about the demo, together with additional screen shots, are available at: `http://secreta.uop.gr/`.

10.5441/002/edbt.2014.58

not support algorithms for transaction data, nor methods for anonymizing *RT*-datasets. Moreover, none of these systems employs utility requirements. We believe that the distinctive features of SECRETA can greatly assist data publishers in making informed decisions on publishing anonymized data.

2. OVERVIEW OF SECRETA

This section describes the components of our system, which we broadly divide into *frontend* and *backend* components. The frontend offers a Graphical User Interface (GUI), which enables users to: (a) issue anonymization requests, and (b) visualize and store experimental results. The backend consists of components for servicing anonymization requests and for conducting experimental evaluations. The architecture of SECRETA is presented in Figure 1.

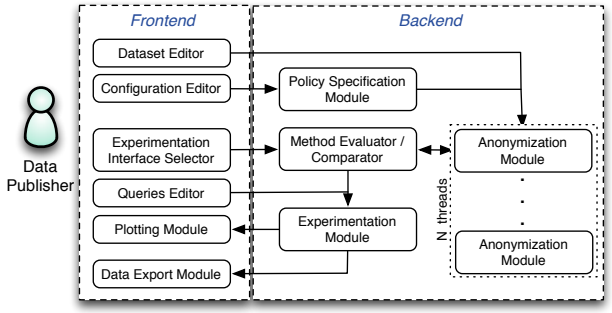

Figure 1: Architecture of SECRETA

2.1 Frontend of SECRETA

The frontend is implemented using the QT framework (`https://qt-project.org`). Using the provided GUI, users can: (a) select datasets for anonymization, (b) specify hierarchies and query workloads, (c) select and configure anonymization algorithms, (d) execute experiments and visualize the experimental results, and (e) export anonymized datasets and experimental results, in a variety of formats. In what follows, we detail the components of the frontend.

Dataset Editor: It enables users to select datasets for anonymization. The datasets can have relational and/or transaction attributes, and they need to be provided in a Comma-Separated Values (CSV) format. Once a dataset is loaded to the Dataset Editor, the user can modify it (edit attribute names and values, add/delete rows and attributes, etc.) and store the changes. The user can also generate data visualizations, such as histograms of attributes. Figure 2 shows a loaded dataset and some visualizations.

Configuration Editor: It allows users to select hierarchies and to specify utility and privacy policies. Hierarchies are used by all anonymization algorithms, except COAT [7] and PCTA [5], whereas utility and privacy policies are only used by these two algorithms to model such requirements. Hierarchies and policies can be uploaded from a file, or automatically derived from the data, using the algorithms in [7].

Queries Editor: This component allows specifying query workloads, which will be used to evaluate the utility of anonymized data in query answering. The system supports the same type of queries as [12], and uses Average Relative Error (ARE) [12] as a defacto utility indicator. The query workloads can be loaded from a file and edited by the user, or be inserted directly using the GUI (see Figure 2).

Experimentation Interface Selector: This component selects the operation mode of SECRETA. Figure 3 shows an interface of the Evaluation mode, in which users can evaluate a given algorithm, while Figure 4 shows an interface of the Comparison mode, which allows users to compare multiple algorithms. Through these interfaces, users can select and configure the algorithm(s) to obtain the anonymized data, store the anonymized dataset(s), and generate visualizations that present the performance of the algorithm(s).

Plotting Module: This module is based on the QWT library (`http://qwt.sourceforge.net/`) and supports a series of data visualizations that help users analyze their data and understand the performance of anonymization algorithms, when they are applied with different configuration settings. Specifically, users can visualize information about: (a) *the original/anonymized dataset* (e.g., histograms of attributes, relative difference of the frequency between an original and a generalized value), and (b) *anonymization results*, for *single* and *varying* parameter execution. In single parameter execution, the results are derived with fixed, user-specified parameters and include frequencies of generalized values in relational or set-valued attributes, runtime, etc. In varying parameter execution, the user selects the start/end values and step of a parameter that varies, as well as fixed values for other parameters. The plotted results include data utility indicators and runtime vs. the varying parameter.

Data Export Module: This module allows exporting datasets, hierarchies, policies, and query workloads, in CSV format, and graphs, in PDF, JPG, BMP or PNG format.

2.2 Backend of SECRETA

The backend of our system is implemented in C++. For each mode of operation, SECRETA invokes one or more instances of the Anonymization Module with the specified algorithm and parameters. The anonymization results are collected by the Method Evaluator/Comparator component and forwarded to the Experimentation Module. From there, results are forwarded to the Plotting Module, for visualization, and/or to the Data Export Module, for data export.

Policy Specification Module: This module invokes algorithms that automatically generate hierarchies [10], as well as the strategies in [7], which generate privacy and utility policies. The hierarchies and/or policies are used by the Anonymization Module (to be described later).

Method Evaluator/Comparator: This component implements the functionality that is necessary for supporting the interfaces of the Evaluation and of the Comparison mode. Based on the selected interface, anonymization algorithm(s) and parameters, this component invokes one or more instances (threads) of the Anonymization Module. After all instances finish, the Method Evaluator/Comparator component collects the anonymization results and forwards them to the Experimentation Module.

Anonymization Module: This component is responsible for executing an anonymization algorithm with the specified configuration. SECRETA supports 9 algorithms; 4 of them are applicable to datasets with relational atrtributes (Incog-

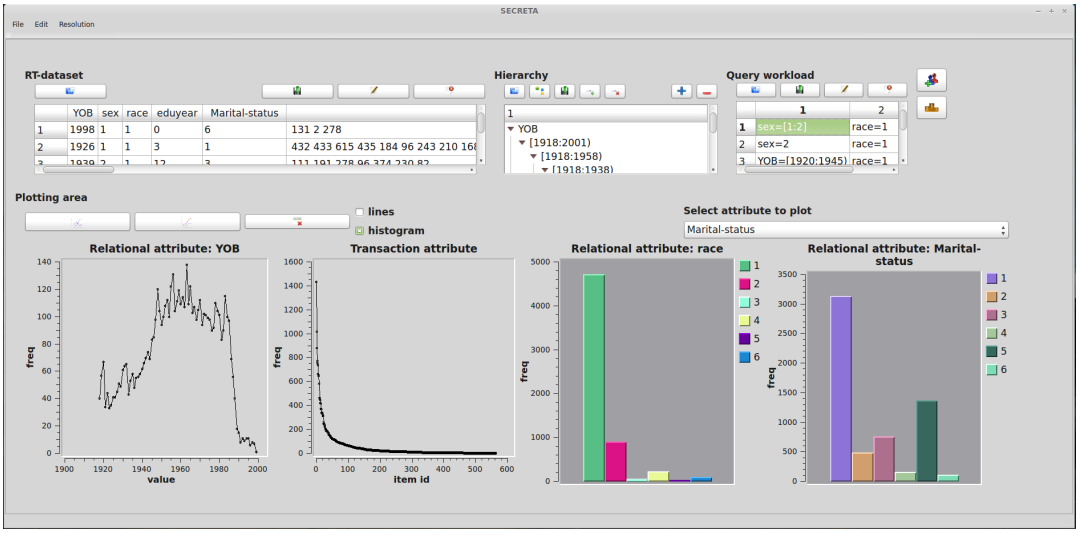

Figure 2: Main screen of SECRETA

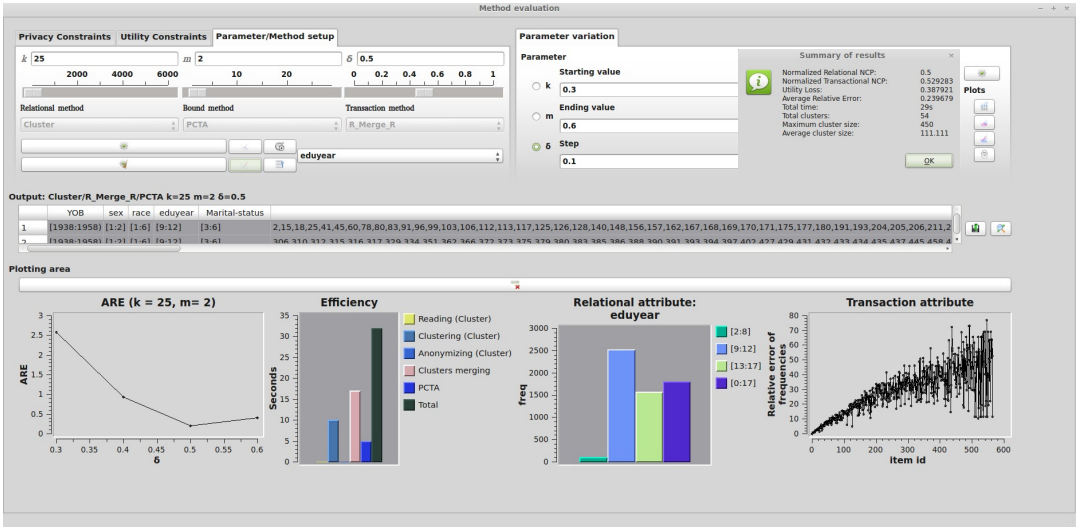

Figure 3: Evaluation mode: Method evaluation screen of SECRETA

nito [6], Cluster [9], Top-down [4], and Full subtree bottom-up), and 5 to datasets with transaction attributes (COAT [7], PCTA [5], Apriori, LRA and VPA [10]). Additionally, it supports 3 bounding methods (**R**MERGE$_r$, **T**MERGE$_r$, **RT**MERGE$_r$) [9], which enable the anonymization of *RT*-datasets by combining two algorithms, each designed for a different attribute type (e.g., Incognito and COAT).

Experimentation Module: This module is responsible for producing visualizations of the anonymization results and of the performance of the anonymization algorithm(s), in the case of *single* and *varying* parameter execution. For visualizations involving the computation of ARE, input is used from the Queries Editor module. The produced visualizations are presented to the user, through the Plotting Module, and can be stored to disk, using the Data Export module.

3. DEMONSTRATION PLAN

During the demonstration, attendees will be able to use SECRETA to: (a) create, edit and analyze a dataset, and (b) execute two different scenarios that demonstrate the modes, functionality range, and potential of the system.

Using the Dataset Editor: The demonstration will start by allowing the user to load a ready-to-use *RT*-dataset. After that, the user will be able to edit the attribute names of the dataset, as well as the values in some records. These operations can be performed directly from the input area (top-left pane in Figure 2), and the user may overwrite the existing dataset with a modified one, or export it to a file. Subsequently, the user will analyze the dataset by plotting histograms of the frequency of values in any attribute (bottom pane in Figure 2).

Using the Configuration and Queries Editor: The user will load a predefined hierarchy from a file. This hierarchy is fully browsable and editable, through the hierarchy area (top-mid pane in Figure 2). Then, the user will

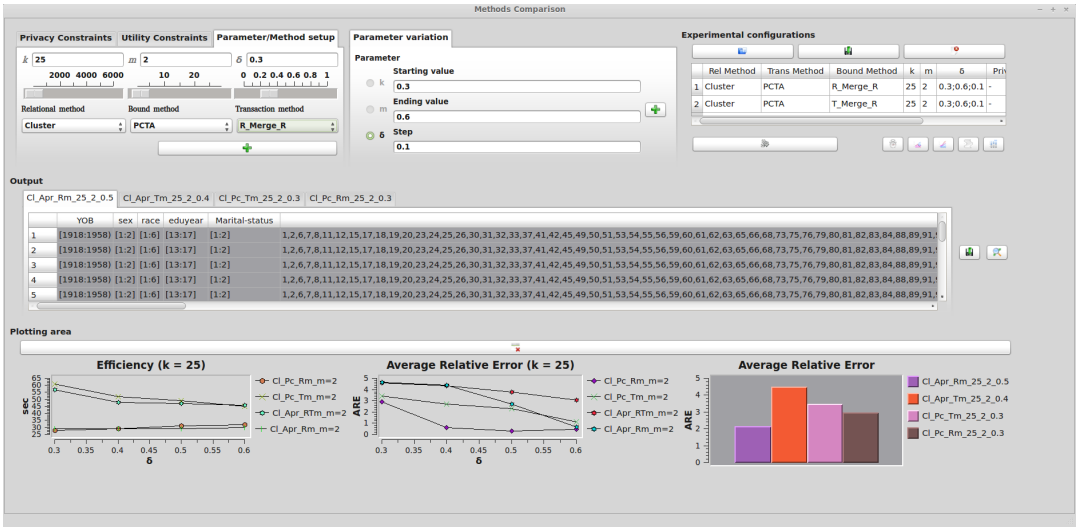

Figure 4: Comparison mode: Methods comparison screen of SECRETA

load a preconstructed query workload from a file, edit the query values using the query workload area (top-right pane in Figure 2), and follow either of the two following scenarios.

Evaluating a method for *RT*-datasets: In this scenario, the users will configure, apply, and evaluate a method, in a series of steps. First, they will use the "Method evaluation" interface (Figure 3) and set the values for parameters k, m, δ, by inputting them directly in the form, or by using the corresponding slider (top-left pane in Figure 3). Then, they may select two algorithms, one for anonymizing the relational attributes, and one for the transaction attribute, and a bounding method for combining the selected algorithms.

Next, the users will initiate the anonymization process. When this process ends, a message box with a summary of results will be presented and the anonymized dataset will be displayed in the output area (middle pane in Figure 3). Last, the users will select a number of data visualizations. These visualizations will be presented in the plotting area (bottom pane in Figure 3) and may illustrate any combination of the following: (a) ARE scores for various parameters (e.g., for varying δ and fixed k and m), (b) the time needed to execute the algorithm and its different phases, (c) the frequency of all generalized values, in a selected relational attribute, and (d) the relative error between the frequency of the transaction attribute values, in the original and the anonymized dataset.

Comparing methods for *RT*-datasets: In this scenario, the users will compare multiple anonymization methods. Using the "Methods comparison" interface (shown in Figure 4), they will: (a) select algorithms for anonymizing each type of attributes, as well as a bounding method, (b) set the values for parameters that will be fixed, as described above (top-left pane in Figure 4), and (c) choose a varying parameter (top-mid pane in Figure 4), along with its start/end value and step. The choices for (a) to (c) comprise a configuration, which will be added into the experimenter area (top-right pane in Figure 4). Similar configurations will be created by the users for at least another method. After the methods are applied, the users will select various graphs, which will be displayed in the plotting area (bottom pane in Figure 4).

4. CONCLUSION

In this paper, we presented SECRETA, a system that helps data publishers analyze the performance of anonymization algorithms and make informed decisions on publishing anonymized data. Our system allows evaluating and comparing a range of different algorithms, in an interactive and progressing way. In the future, we will extend our system, by incorporating additional algorithms, such as those in [2].

Acknowledgements

G. Poulis is supported by Heraclitus II, S. Skiadopoulos by EICOS/Thalis, and G. Loukides by a Research Fellowship from the Royal Academy of Engineering.

5. REFERENCES

[1] `http://www-03.ibm.com/software/products/us/en/infosphere-optim-data-privacy/`.

[2] J. Cao, P. Karras, C. Raïssi, and K. Tan. *rho*-uncertainty: Inference-proof transaction anonymization. *PVLDB*, 3(1):1033–1044, 2010.

[3] C. Dai, G. Ghinita, E. Bertino, J.-W. Byun, and N. Li. Tiamat: a tool for interactive analysis of microdata anonymization techniques. *PVLDB*, 2(2), 2009.

[4] B. Fung, K. Wang, and P. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, 2005.

[5] A. Gkoulalas-Divanis and G. Loukides. Utility-guided clustering-based transaction data anonymization. *TDP*, 5(1):223–251, 2012.

[6] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: efficient full-domain k-anonymity. In *SIGMOD*, 2005.

[7] G. Loukides, A. Gkoulalas-Divanis, and B. Malin. COAT: Constraint-based anonymization of transactions. *Knowledge and Information Systems*, 28(2):251–282, 2011.

[8] National Institutes of Health, 2013. Data repositories. `http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html`.

[9] G. Poulis, G. Loukides, A. Gkoulalas-Divanis, and S. Skiadopoulos. Anonymizing data with relational and transaction attributes. In *ECML/PKDD*, 2013.

[10] M. Terrovitis, N. Mamoulis, and P. Kalnis. Local and global recoding methods for anonymizing set-valued data. *VLDB J.*, 20(1):83–106, 2011.

[11] X. Xiao, G. Wang, and J. Gehrke. Interactive anonymization of sensitive data. In *SIGMOD*, 2009.

[12] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.-C. Fu. Utility-based anonymization using local recoding. In *KDD*, pages 785–790, 2006.