TIAMAT: a Tool for Interactive Analysis of Microdata Anonymization Techniques

Chenyun Dai[1], Gabriel Ghinita[1], Elisa Bertino[1], Ji-Won Byun[*] [2], Ninghui Li[1]

[1]Dept. of Computer Science
Purdue University
West Lafayette, IN 47907
{daic, gghinita, bertino, ninghui}@cs.purdue.edu

[2]Database Security
Oracle Corp.
Redwood City, CA 94065
ji-won.byun@oracle.com

ABSTRACT

Releasing detailed data (*microdata*) about individuals poses a privacy threat, due to the presence of *quasi-identifier (QID)* attributes such as age or zip code. Several privacy paradigms have been proposed that preserve privacy by placing constraints on the value of released QIDs. However, in order to enforce these paradigms, data publishers need tools to assist them in selecting a suitable anonymization method and choosing the right system parameters. We developed *TIAMAT*, a tool for analysis of anonymization techniques which allows data publishers to assess the accuracy and overhead of existing anonymization techniques. The tool performs interactive, head-to-head comparison of anonymization techniques, as well as QID change-impact analysis. Other features include collection of attribute statistics, support for multiple information loss metrics and compatibility with commercial database engines.

1. INTRODUCTION

The problem of privacy-preserving data publication has been intensively studied in recent years. Consider the example of a hospital which gathers large amounts of detailed data (*microdata*) about patients. Such data can be mined in order to derive certain disease patterns, and can benefit medical research. However, releasing the microdata introduces a privacy threat: even after removing directly identifying information, such as name or SSN, the data still contain *quasi-identifier (QID)* attributes (e.g., age, zipcode) that can help an attacker learn the identity of individuals whose personal information is included in the microdata.

To prevent disclosure of sensitive information, the k-anonymity paradigm [8, 9] requires each published record to be indistinguishable with respect to the QID attribute set among an *anonymized group* of $k-1$ other records. k-anonymity is achieved through QID *generalization*, which replaces exact values with value ranges. If a *Value Generalization Hierarchy (VGH)* [4] exists, a leaf-node value is replaced with one of its ancestor nodes. Inherently, generalization introduces information loss, which must be minimized in order to preserve data accuracy.

In the quest for a good privacy-accuracy trade-off, data publishers are faced with several important decisions, such as selecting a suitable k-anonymization algorithm, choosing which QID attributes to release (and their associated VGHs), and determining an appropriate value of k. Since each dataset has its particular characteristics, no single set of parameters can accommodate all scenarios, and finding the right settings is a cumbersome and error-prone process.

We propose TIAMAT: a Tool for Interactive Analysis of Microdata Anonymization Techniques. TIAMAT focuses on k-anonymity, but it can be easily extended to support other privacy-preserving paradigms that employ QID generalization, such as ℓ-diversity [7] and t-closeness [6]. The tool allows data publishers to analyze the accuracy and runtime performance of various k-anonymization techniques, and to find suitable parameter settings for anonymization. TIAMAT supports two analysis types:

- *Comparative Analysis of Anonymization Techniques:* TIAMAT can be used in conjunction with any QID-generalization based anonymization solution, and supports head-to-head performance comparison of competitor techniques. Multiple information loss metrics are supported, and the results are presented within an interactive visualization framework.
- *QID Change-Impact Analysis:* the tool assists publishers in choosing a suitable set of QID attributes. To this extent, TIAMAT provides a powerful function that anonymizes data with respect to all combinations of QID chosen from a user-specified attribute set.

TIAMAT includes several other features that facilitate the efficient analysis of anonymization techniques:

- *VGH Editing:* users can inspect and edit VGHs in a visual manner. VGHs can also be imported from and exported to external files.
- *Attribute Statistics Collection:* users can collect and visualize statistics on attribute value distributions; the results can be visualized in the form of histograms.
- *Integration with DB Engines:* the microdata can be retrieved directly from any SQL-compliant database

[*]Ji-Won Byun worked on this project while he was a student at Purdue University.

VLDB '09, August 24-28, 2009, Lyon, France

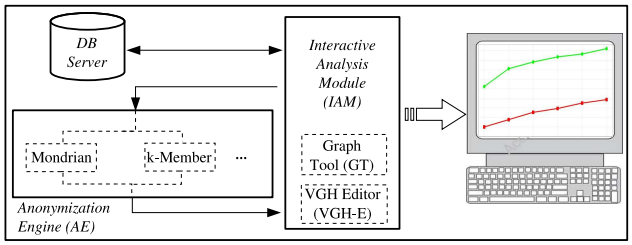


Figure 1: System Architecture

server. Furthermore, the user has the option to export the anonymized data back to the database.

2. SYSTEM ARCHITECTURE

The TIAMAT prototype is developed on top of Java SDK 6. Its architecture is outlined in Figure 1: the user interacts directly with the *Interactive Analysis Module (IAM)*, which handles user input and visualization of results. The data is retrieved from an SQL database server (Oracle 11G). *IAM* includes two sub-components: the *Value Generalization Hierarchy Editor (VGH-E)*, which allows users to view and edit VGHs, and the *Graph Tool (GT)*, a result plotting tool based on the JFreeChart[1] library.

IAM sends data anonymization requests to the *Anonymization Engine (AE)* module, which performs QID generalization. *AE* has a modular design with a generic interface that facilitates the incorporation of any generalization-based anonymization method. Currently, *AE* supports two algorithms: Mondrian [5] and k-Member [1]. The former relies on partitioning of the quasi-identifier space, whereas the latter performs clustering of records.

Information loss (i.e., data accuracy) can be evaluated with respect to two metrics: *Global Certainty Penalty (GCP)* [3] and *Classification Metric (CM)* [4, 1]. GCP measures the amount of distortion introduced by generalization, whereas CM quantifies the classification homogeneity with respect to some label attribute within each anonymized group. Both metrics have values in the $[0, 1]$ range, where 0 is the ideal value and 1 signifies complete information loss.

3. DEMONSTRATION OVERVIEW

This section gives an overview of the TIAMAT features that will be presented during the demonstration. We anonymize the *Adult* dataset (30,162 records) from the UC Irvine repository[2], consisting of 6 numerical and 9 categorical attributes. The dataset has been used extensively in previous work [4, 5, 1, 3] to evaluate k-anonymization techniques. In Section 3.1, we discuss features that help users prepare the dataset for anonymization, and compute dataset statistics. Next, in Section 3.2, we show how TIAMAT allows users to decide which anonymization method is more suitable for their data, by comparing head-to-head the results obtained by alternative techniques. Later, in Section 3.3 we show how TIAMAT assists users to identify attributes that cause excessive information loss when they are included in the quasi-identifier set.

[1]http://www.jfree.org/jfreechart/
[2]http://www.ics.uci.edu/~mlearn/MLRespository.html

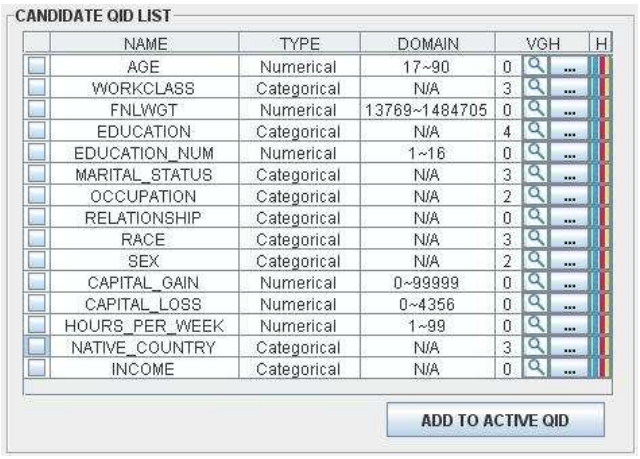


Figure 2: Quasi-Identifier Attributes Panel for *Adult* Dataset

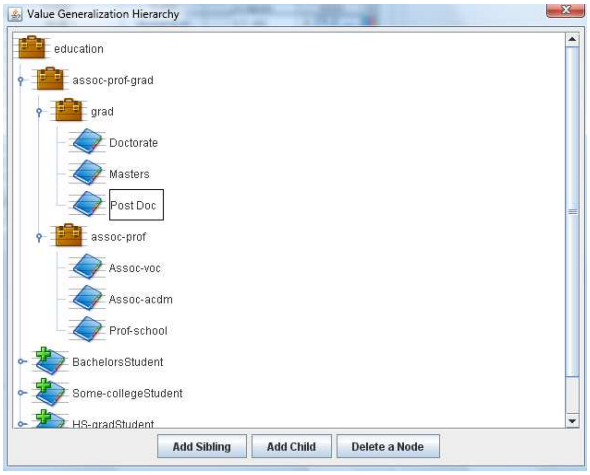


Figure 3: Customizing the *Education* Attribute Taxonomy Tree with the VGH Editor (*VGH-E*) Tool

3.1 Dataset Preparation and Statistics

TIAMAT inter-connects with SQL-compliant databases where microdata tables are stored, and automatically populates the candidate QID list from the table schema (Figure 2). Information about the type and domain of each attribute is included in the QID list. In addition, the number of levels in the taxonomy (i.e., VGH) tree associated to each categorical attribute is shown. The users can visualize and edit the attribute VGH using the VGH-E tool. Figure 3 shows the VGH for attribute *Education*. Users can change the VGH in an interactive manner, or can import a hierarchy from an external file. In this example, a new node (*Post-Doc*) is added to the taxonomy tree.

TIAMAT assists users in deciding what QID attributes to release, by providing statistical information about attribute values. Characteristics such as attribute domain range, mean and median values, and value distribution typically influence the information loss incurred by generalization. A histogram with the frequency of attribute value

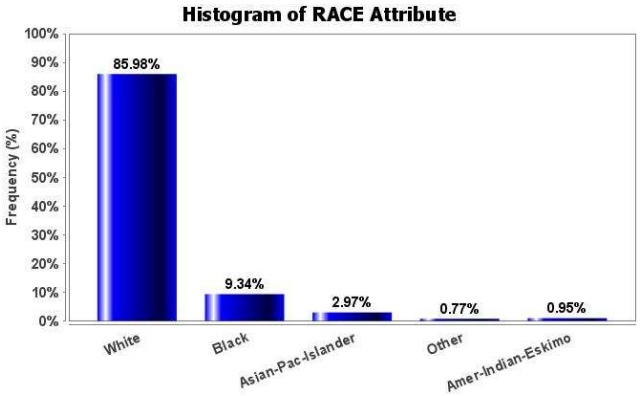


Figure 4: Histogram Showing the Value Distribution for the *Race* Attribute

occurrences can be visualized by clicking the attribute's H button in the QID panel. Figure 4 shows an example of collected statistics for attribute *Race*. Such information may be useful, for instance, to adjust the VGH tree. In this example, the user learns that the $Asian - Pac - Islander$ and $Amer - Indian - Eskimo$ values are very sparse in the dataset. Therefore, it is very likely that in groups containing such records, the *Race* attribute will be generalized. The user may decide to place all such low-frequency values close to each other in the VGH tree, such that k-anonymization would include them in the same groups, instead of combining them with high-frequency values. This way, groups containing high-frequency *Race* values can be homogeneous, and no generalization is necessary, leading to lower information loss.

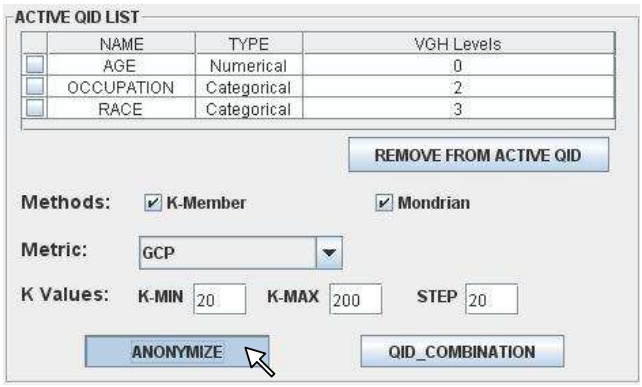
3.2 Head-to-head Comparison Analysis

In this demonstration scenario, alternative k-anonymization techniques are evaluated head-to-head with respect to data accuracy and execution time. The evaluation is performed for a fixed QID attribute combination, chosen by the user. In Section 3.3, we show how TIAMAT assists users in choosing a suitable QID for publication.

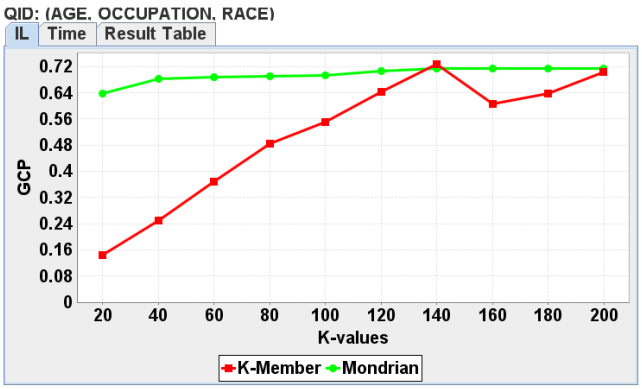
Data publishers may wish to consider several distinct k-anonymization techniques in order to determine which one is more suitable for their data. TIAMAT allows direct comparison between competitor algorithms. Consider an example where attributes *Age*, *Occupation* and *Race* are chosen as QID (Figure 5a). The user wishes to compare Mondrian and k-Member with respect to the GCP metric[3]. This submission screen also allows the user to select the range of k (in this case $20 - 200$), as well as the interval between consecutive k values.

Figure 5b shows the anonymization results. There are three separate tabs. The *IL* and *Time* tabs present summary plots of information loss (GCP) and execution time, respectively (the time measurement is omitted from the figure). Based on the variation of information loss with k, the user can decide what is a suitable value of k to choose, in order to obtain a good trade-off between privacy and accuracy. The *Result Table* tab allows the user to inspect the

[3]The head-to-head comparison scenario can be applied in a similar manner with the classification metric.



(a) Choosing Anonymization Parameters



(b) Information Loss Visualization Panel

AGE	WORKCLASS	FNLWGT	EDUCATION	MARITAL_STATUS	OCCUPATION	RELATIONSHIP
25.0~36.0	Private	211265.0	Some-college	other	Occupation	Other-relative
25.0~36.0	Private	189775.0	Some-college	other	Occupation	Own-child
25.0~36.0	Private	212563.0	Some-college	other	Occupation	Unmarried
25.0~36.0	Private	70282.0	Some-college	other	Occupation	Unmarried
28.0~47.0	Private	88419.0	education	Never-married	Exec-managerial	Not-in-family
28.0~47.0	Private	163003.0	education	Never-married	Exec-managerial	Other-relative
28.0~47.0	Private	348491.0	education	Never-married	Exec-managerial	Not-in-family
28.0~47.0	Private	200734.0	education	Never-married	Exec-managerial	Unmarried
29.0~43.0	Private	419721.0	HS-grad	Never-married	Other-service	Unmarried
29.0~43.0	Private	206365.0	HS-grad	Never-married	Other-service	Not-in-family
29.0~43.0	Private	39581.0	HS-grad	Never-married	Other-service	Not-in-family
29.0~43.0	Private	70240.0	HS-grad	Never-married	Other-service	Own-child

(c) Anonymized Table (k-Member)

Figure 5: Head-to-head Comparison Analysis

contents of the detailed anonymized tables for each value of k, as shown in Figure 5c (different anonymized groups are highlighted with distinct background colors). The user has the option to store the anonymized tables back to the database server.

In this example, k-Member is superior to Mondrian in terms of information loss for most of the k value range. However, the GCP exhibits a pronounced spike for anonymity degree $k = 140$. The user may wish to remove the cause for this increase in GCP which deteriorates data accuracy. Specifically, s/he may want to determine which of the QID attributes have the most significant contribution to the increase in GCP. Next, we show how TIAMAT can assist the user in this task.

3.3 QID Change-Impact Analysis

In general, the choice of QID attributes has a major impact in the information loss incurred by anonymization: in

some cases, the inclusion of a certain attribute may seriously decrease data accuracy. In particular, attributes associated with VGHs are more likely to determine spikes in information loss when the values of other parameters change (e.g., k or number of records). The cause of such behavior is the inclusion in the same anonymized group of attribute values from distant VGH sub-trees. It is important for data publishers to be able to evaluate several QID attribute sets, and choose for publication the most suitable one.

TIAMAT allows users to automate the QID selection process. Continuing the previous example, the user can choose the *QID_Combination* feature (from the screen in Figure 5a), which generates all possible combinations of attributes from the active QID list, and anonymizes the data with respect to each combination. Figure 6 shows the outcome of the QID change-impact analysis (for brevity, we include only 6 out of all $2^3 - 1 = 7$ combinations of one, two or three attributes from the candidate QID list. The user can observe that in most of the cases where attribute *Occupation* is included, the spikes in GCP are present. Furthermore, the (*Age*, *Race*) combination which does not include *Occupation* presents a smooth trend. Hence, the user can conclude that the *Occupation* attribute is the one causing the spike in information loss. If *Occupation* is not an essential part of the data, the user may wish to remove it completely, since it may affect the accuracy of the other generalized attributes. Alternatively, if the attribute is important, the user can employ the VGH-E tool, and re-organize the VGH associated with attribute *Occupation*. Finally, the user can decide to improve the accuracy on attribute *Occupation* by assigning it a higher weight in the anonymization process (for details on weighted information loss metrics, please see reference [3]). Therefore, using the change-impact analysis tool, the user can refine the QID attribute set that is finally chosen for publication, and improve the accuracy of released data.

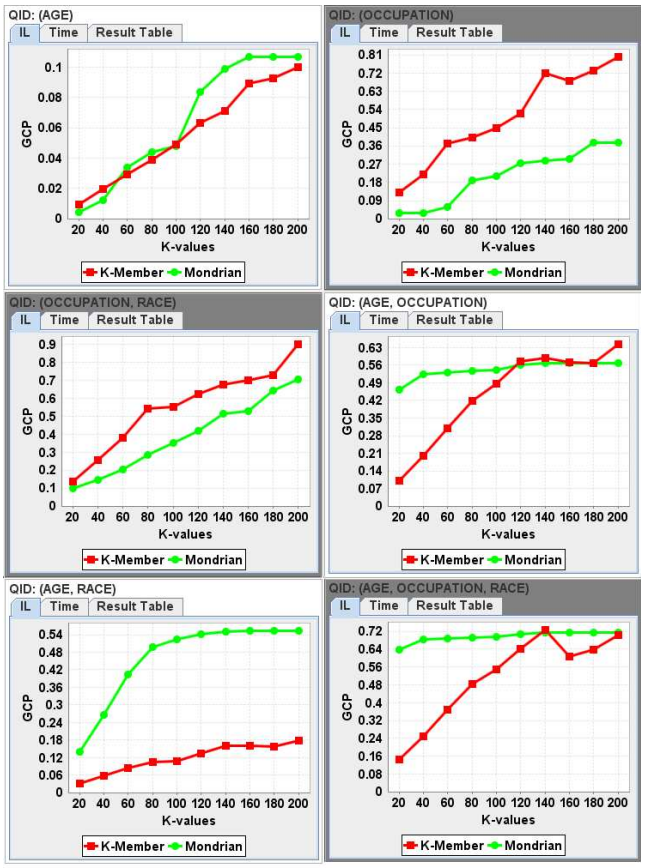


Figure 6: QID Change-Impact Analysis

4. CONCLUSIONS

TIAMAT is a visual tool that helps data publishers select a suitable k-anonymization transformation and its corresponding parameters in order to protect their data. In future work, we plan to augment the tool with a processing engine for aggregate queries and classification tasks on top of the generalized data. This allows users to optimize the anonymization process for specific tasks. We also plan to include support for releasing multiple anonymized versions of the same microdata table ([2, 10]). Furthermore, we will extend the set of available k-anonymization techniques, by adding support for algorithms such as [3, 11]. Another interesting feature that we will consider is to allow automated random sampling of data (essential for large datasets and/or slower k-anonymization algorithms).

5. ACKNOWLEDGMENTS

The work by E. Bertino, C. Dai, G. Ghinita and N. Li was partially supported by a Google grant for research on "Utility and Privacy in Data Anonymization"

6. REFERENCES

[1] J.-W. Byun, A. Kamra, E. Bertino, and N. Li. Efficient k -Anonymization Using Clustering Techniques. In *In Proc. of DASFAA*, pages 188–200, 2007.

[2] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li. Secure Anonymization for Incremental Datasets. In *Third VLDB Workshop on Secure Data Management (SDM'06)*, 2006.

[3] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis. Fast Data Anonymization with Low Information Loss. In *Proc. of VLDB*, pages 758–769, 2007.

[4] V. S. Iyengar. Transforming Data to Satisfy Privacy Constraints. In *Proc. of SIGKDD*, pages 279–288, 2002.

[5] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian Multidimensional k-Anonymity. In *Proc. of ICDE*, 2006.

[6] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *Proc. of ICDE*, pages 106–115, 2007.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-Diversity: Privacy Beyond k-Anonymity. In *Proc. of ICDE*, 2006.

[8] P. Samarati. Protecting Respondents' Identities in Microdata Release. *IEEE TKDE*, 13(6):1010–1027, 2001.

[9] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[10] Y. Tao and X. Xiao. m-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets. In *Proc. of ACM SIGMOD*, pages 689–700, 2007.

[11] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu. Utility-Based Anonymization Using Local Recoding. In *Proc. of SIGKDD*, pages 785–790, 2006.