CAT: The Cornell Anonymization Toolkit

Chapter 1

Introduction

The Cornell Anonymization Toolkit (CAT) is designed for interactively anonymizing published dataset to limit identification disclosure of records under various attacker models. It was developed by the Computer Science Department at Cornell University [5]. The toolkit contains the following functions:

1. **Data Generalization,** anonymizing data using generalization [4], which transforms non-sensitive attribute values in the data into value ranges, so as to prevent an adversary from identifying individuals by linking these attributes with public available information.

2. **Risk Analysis,** evaluating the disclosure risks of each record in anonymized data based on user-specified assumptions about the adversary's background knowledge. In addition, the distribution of the disclosure risks of all records in the dataset can be illustrated in a histogram.

3. **Utility Evaluation,** comparing contingency tables and density graphs between the original and anonymized data. Both measurements provide users an intuitive way to learn the statistical distortion incurred in the dataset through anonymization.

4. **Sensitive Record Manipulation,** applying special treatment to records with much higher disclosure risks than most other records. Such records could be the outliers in the dataset, and they may severely degrade the quality of anonymization. Users can then eliminate these sensitive records with high disclosure risks.

5. **Visualization and Interaction.** The result dataset of the above functions can be shown in the user interface (see Figure 1.1), and users can apply the above process iteratively and observe the results until they obtain a satisfactory anonymization in both terms of data privacy and utility.

The dataset to be anonymized is kept in main memory, and all the above functions are executed against this main-memory resident data, which is not output to disk until the users are satisfied with the anonymization result.

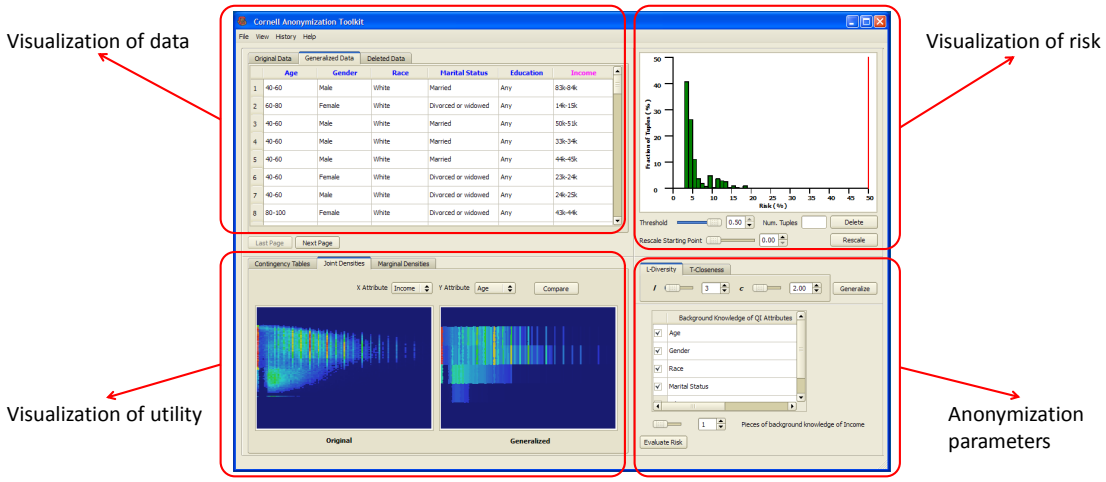

Figure 1.1: CAT User Interface

Chapter 2

Background Knowledge and CAT Overview

Data privacy is an important issue when one wants to make use of data that involve individuals' sensitive information, such as data released by hospitals, government agencies, insurance companies, etc. For example, suppose a dataset of medical records is released for public use. Even though identifier attributes such as names, social security numbers, or phone numbers are not contained. Adversary can still identify certain individual's medical records in the data since other attributes can be combined to act as an "identifier" to the individual: it is not common for two individuals to have the same birth date, and even less common for them to also have the same gender, or live in the same zip code. Hence such attributes are called *quasi-identifiers*. By linking them with external data, the *sensitive attributes* of the individuals, here the medical status, are revealed. Therefore the individual privacy is disclosed by releasing this dataset. Research on protecting the privacy of individuals has received contributions from many fields, including computer science and statistics. For a thorough survey of research work in privacy-preserving data publishing, read [1].

2.1 Privacy Criterion

Given a anonymization result, the privacy criterion defines whether the result is safe for release or not. A large body of work has provided numerous criteria of privacy, which characterized the assumptions about the adversary's background knowledge used for attacking individual privacy. CAT have implemented two privacy criteria: l-Diversity [3] and t-Closeness [2]. Given a certain privacy criterion, there are a number of anonymization strategies to achieve this criterion, such as data generalization, data swapping, data data perturbation, etc. CAT currently only support the data generalization mechanism, which will be introduced in the next section.

2.2 Data Generalization

Data generalization is a strategy for protecting individual privacy in released microdata records [4]. The idea is that by reconstructing a more "general" and semantically consistent domain for the attributes and transforming it values to this domain, it would be much harder to identify individuals by linking this attribute with external data. For example, the domain of attribute Country can be generalized by replacing country values with continents.

Consider a dataset consisting of a single table of microdata records having categorical and numeric attributes. For unordered categorical attributes, this idea of generalization can be implemented through the user-defined *generalization hierarchies*. For numeric attributes, generalization can instead take the form of a coarsened range of values.

2.3 Disclosure Risk

Disclosure risk is a term frequently used in the statistics literature to refer to quantifiable estimates of the possibility of a privacy disclosure. Measuring disclosure risk is a key step in defining privacy criteria, since it is also based on the assumptions about the adversary's background knowledge. For example, l-diversity model assumes that the adversary may have information of the nonsensitive 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 a income of 10k." Then disclosure risk of an individual is quantified as the adversary's posterior probability of inferring the correct values of its sensitive attributes, after combing the anonymized data with the background knowledge.

2.4 Utility Metrics

Besides individual privacy guarantees, data publishers are also concerned with the amount of useful information reside in the anonymized data. Therefore measurements are needed to evaluate the utility of different anonymization candidates in order for release. Possible measurements include Loss Metric, Classification Metric, KL-Divergence, L_p Norm, etc. Instead of implementing these metrics, CAT illustrates the information loss by comparing contingency tables and density graphs between the original and generalized data. That is because by comparing contingency tables and density graphs on the graphical interface, users can get a more intuitive understanding about the first- and second-order statistical distortion incurred through the anonymization.

Chapter 3

Using CAT

3.1 Data Files

CAT requires the input data files to be separated into microdata and metadata. The metadata file encodes the attribute values, and hence the microdata file only needs to store the attribute codes, largely reducing the size of the file. An example metadata file and the corresponding microdata file is shown below.

- **Metadata File**:

```
5   300000   // Size of the dataset

127 0   0   Age   // First attribute is integerical with domain size 127
0   126

2   0   1   Gender   // Second attribute is categorical with domain size 2
1   Male
2   Female

9   0   1   Race   // Third attribute is categorical with domain size 9
1   White
2   Black
3   American Indian or Alaska Native
4   Chinese
5   Japanese
6   Other Asian or Pacific Islander
7   Other race, nec
8   Two major races
9   Three or more major races
```

```
18  0   1   Education   // Fourth attribute is categorical with domain size 18
00  Not applicable
01  No school completed
02  Nursery school
03  Kindergarten
04  1st-4th grade
05  5th-8th grade
06  9th grade
07  10th grade
08  11th grade
09  12th grade, no diploma
10  High school graduate, or GED
11  Some college, no degree
12  Associate degree, occupational program
13  Associate degree, academic program
14  Bachelor's degree
15  Master's degree
16  Professional degree
17  Doctorate degree

100 0   1   Income   // Fifth attribute is categorical with domain size 100
0   0k-1k
1   1k-2k
2   2k-3k
3   3k-4k
4   4k-5k
5   5k-6k
6   6k-7k
7   7k-k8
8   8k-9k
9   9k-10k

...

97  97k-9k8
98  98k-99k
99  99k-100k
```

- **Microdata File**:

```
1   48  1   1   10  83
2   66  2   1   6   14
3   51  1   1   11  50
4   48  1   1   15  33
5   48  1   1   5   44
6   50  2   1   11  23
```

```
7   59  1   1   11  24
8   83  2   1   14  43
9   53  1   2   12  14
10  64  2   1   9   23
```

For metadata file, the first line has two numbers. The former one specify the number of attributes M in the dataset, and the latter one specify the number of records N in the dataset. Hence the size of the dataset is $M \times N$. Then each attribute is defined and single-line-separated. The first line of the definition specify the domain size, the number of missing values, the attribute type (0 for integerical, 1 for categorical, 2 for floating-point), and the attribute name in turn. After that, for categorical attributes, each possible value is encoded; for integerical and floating-point attributes, the range is specified.

As a result, the microdata file stores the data records using only numerical values. The first number is for the record ID, and the second number is for the first attribute, and so on. For example, the third line in the above microdata file actually represents the following record with ID 3:

```
Age Gender Race    Education                Income
51  Male   White   Some college, no degree  50k-51k
```

Although CAT does not literally require that the data attributes be categorical, but treats the numerical attributes as if they were categorical. Therefore besides the dataset itself, CAT also requires users to provide a hierarchy file for each attribute. Note that the leaf level of the hierarchy is already specified in the metadata file. An example hierarchy file for attribute Education of the above metadata is shown below.

```
4   // Number of Levels in the Hierarchy, excluding the leaf level

10  // Number of nodes in the first level
0   3   No primary school
4   4   1st-4th grade
5   5   5th-8th grade
6   9   9th-12th grade
10  10  High school graduate, or GED
11  11  Some college, no degree
12  13  Associate degree
14  14  Bachelor's degree
15  15  Master's degree
16  17  Professional or doctorate degree

5   // Number of nodes in the second level
0   4   4th grade or lower
5   9   5th-12th grade
10  10  High school graduate, or GED
11  11  Some college, no degree
12  17  Associate degree or above
```

```
2   // Number of nodes in the third level
0   9   Below high school
10  17  High school graduate or above

1   // Number of nodes in the fourth level
0   17  Any
```

One can see that the number of internal hierarchy levels is specified firstly in the file. Then each internal level is defined by the number of nodes of the level, and the cover range of the leaf nodes for each node. Note that all the internal level nodes' ranges must completely cover their child level nodes' ranges. For example, given the first level specified above, the second level nodes cannot have a range of $0 - 7$.

3.2 Getting Started

The process of anonymizing a dataset using CAT is illustrated in Figure 3.1. We begin by loading the dataset into CAT.

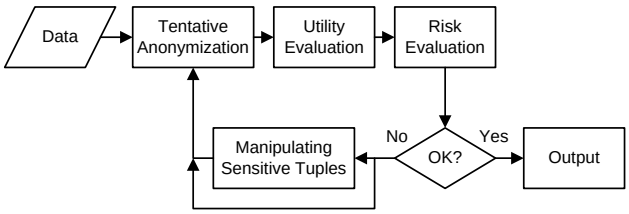

Figure 3.1: Anonymization Process

3.2.1 Read and Initialize Microdata

By clicking on the "File → Load data..", users can specify the input microdata and metadata files (see Figure 3.2). Once loaded, the tuples in the dataset will be shown in the "Original Data" table in the upper-left panel of the user interface. In addition, users need also to choose the quasi-identifiers (QI) and sensitive attribute (SA) by clicking "File → Initialize..". Multiple SAs can be chosen, but only one SA can be selected. From the "Original Data" table, we can see that purple attribute is SA, and the blue attributes are QI. For each QI, users also need to input a hierarchy file (see Figure 3.3).

3.2.2 Generalize Microdata and Analyze Disclosure Risk

Now the dataset is ready to be generalized. Users can choose either the l-Diversity or t-Closeness algorithm to generalize the dataset in the middle-right panel. For example, the l-Diversity widget has two sliders

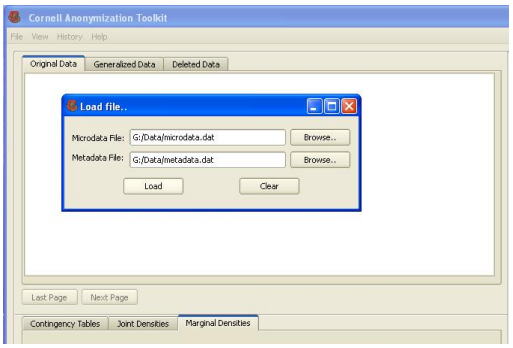

Figure 3.2: Load Dataset File

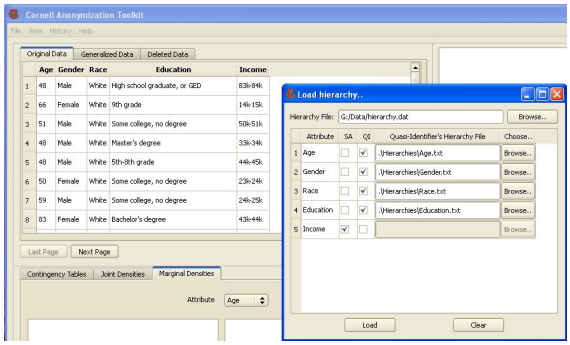

Figure 3.3: Load Hierarchy File

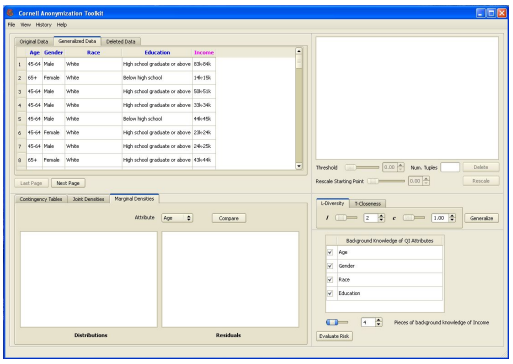

Figure 3.4: Data Generalization

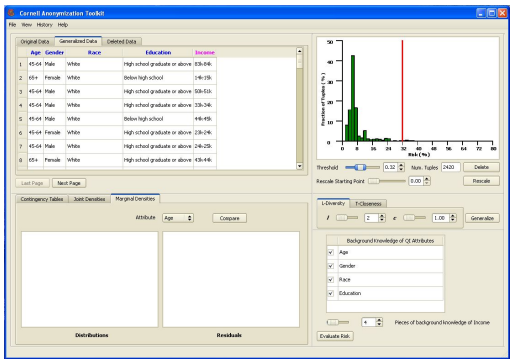

Figure 3.5: Risk Evaluation

that control the two parameters l and c of the algorithm [3]. If users do not have a clear idea of how these parameters should be set, then some initial values can be set (e.g., 2 and 1.0). By clicking the "Generalize" button, the upper-left panel will automatically switch to the "Generalized Data" table showing the anonymized dataset by CAT (see Figure 3.4).

To analyze disclosure risk under the l-Diversity model, users need to choose in the lower-right panel which QIs the adversary is assumed to have the full knowledge of, and the number of pieces of knowledge she knows about the SA. After clicking on the "Evaluate Risk" button, the risk distribution histogram will be shown in the upper-right panel. For example, from Figure 3.5 we can observe that less than 20 percent of the tuples have a risk of 8%.

3.2.3 Evaluate Data Utility

To get an understanding of the utility of an anonymization, users can 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. A contingency table is a table that shows the frequencies for combinations of two attributes. Intuitively, contingency tables show correlations between pairs of attributes. In addition, users can also

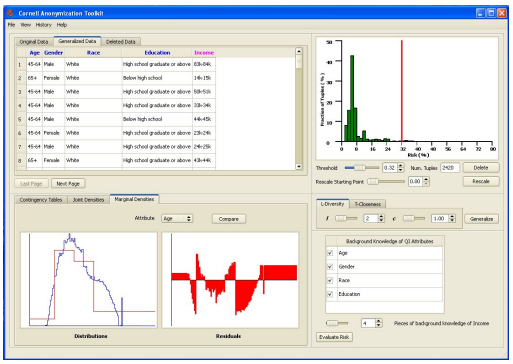

Figure 3.6: Marginal Density Comparison

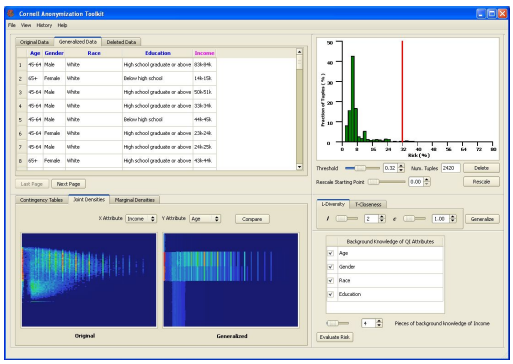

Figure 3.7: Joint Density Comparison

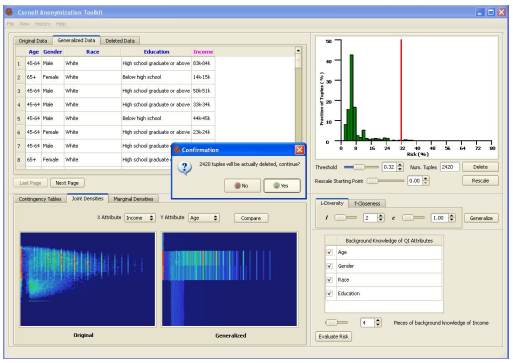

Figure 3.8: Risky Records Deletion

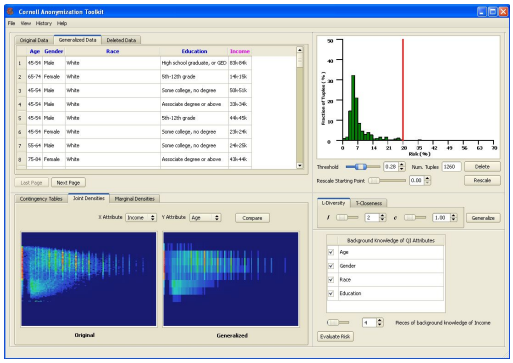

Figure 3.9: Joint Density Closeness Comparison

compare the marginal or joint densities by clicking on the "Marginal Densities" or "Joint Densities" tab. In general, the more similar the graphs are, the more useful information is retained in the anonymized table (see Figure 3.6 and 3.7).

To enhance the anonymization quality, users can eliminate those records with high risk. For example, users can set the "Threshold" slider on the upper-right panel to eliminate records whose risk is higher than the threshold. While the slider moves, CAT will dynamically calculate how many records need to be eliminated. After users confirm the deletion of risky tuples by clicking "Yes", the risk distribution will be recalculated and refreshed on the upper-right panel (see Figure 3.8). After that, users can re-generalize the pruned dataset in order to get a better anonymization. For example, after records with risk higher than 32% are eliminated, the re-generalization can output another anonynimization result with better utility (compare joint density closeness of Figure 3.9 with Figure 3.8).

Furthermore, by clicking on "Histogram → Show Log", users can also undo some generalization and deletion actions from the user log in order to revert to a previous anoanymization version, if they are not satisfied with the current result (see Figure 3.10).

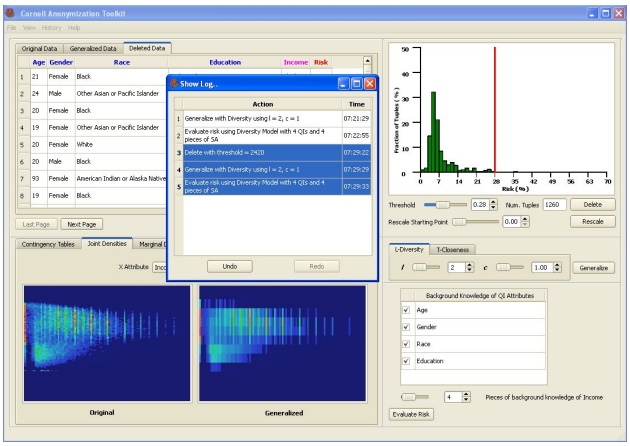

Figure 3.10: Undo/Redo User Log

3.2.4 Publish Anonymized Microdata

Users can apply all the above processes iteratively until they obtain a satisfactory anonymization result. Then the result can be published to the specified output file by clicking on "File → Publish.." (see Figure 3.11). Note that the output file will contain the real attribute values instead of the attribute encodings.

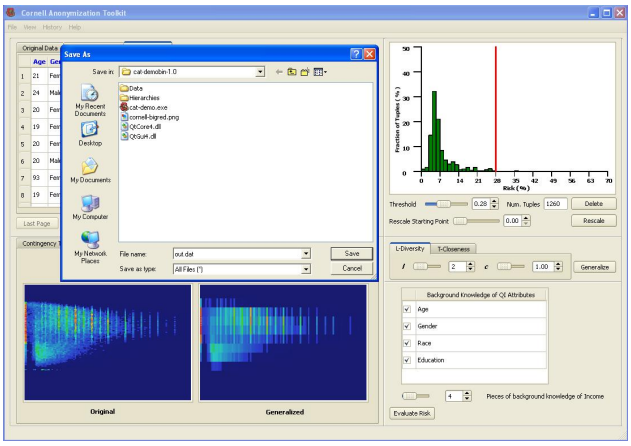

Figure 3.11: Anonymized Data Publication

Chapter 4

Acknowledgements

This material is based upon work supported by the New York State Foundation for Science, Technology, and Innovation under Agreement C050061. Any opinions, findings, conclusions or recommendations expressed are those of the author(s) and do not necessarily reflect the views of NYSTAR. This material is also based upon work supported by the National Science Foundation under Grant 0627680. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

Bibliography

[1] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala. Privacy-preserving data publishing. *Foundations and Trends in Databases*, 2(1-2):1–167, 2009.

[2] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, pages 106–115, 2007.

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

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

[5] X. Xiao, G. Wang, and J. Gehrke. Interactive anonymization of sensitive data. In *SIGMOD Conference*, pages 1051–1054, 2009.