Anonymization Case Study 1:
Randomizing Names and Addresses

The PrivacyAnalytics Tool is being developed as part of a joint research project between the Children's Hospital of Eastern Ontario Research Institute, the University of Ottawa, and Bell Information and Communications Technology Solutions Inc.

Introduction

In this brief case study we will walk the reader through an example of anonymizing a data file using the PrivacyAnalytics tool. PrivacyAnalytics is a desktop data anonymization tool that is being developed as part of a research project between the Children's Hospital of Eastern Ontario Research Institute, the University of Ottawa, and Bell Information and Communications Technology Solutions.

The data set that we start off with contains identifying information: full names and addresses of individuals. We assume that the data set also contains sensitive information pertaining to these individuals. There are a number of scenarios where the disclosure of such a data set could be seen as an invasion of privacy. Example scenarios are:

> **Software Testing Scenario.** A software testing team needs to run a few tests through a health insurance data processing application, and they need real data to make sure that the tests are as realistic as possible. The data that is needed includes names and addresses, as well as financial information about a company's clients. We also assume that the test team is separate from the main business units of the organization. For example, the testing team may be at another site, the testing function has been contracted out to another company, or both of the above with the addition that the testing function was outsourced to a company in India. It would be risky to give the test team real customer data from the production environment.
>
> **Providing Data to Researchers Scenario.** A researcher wants to perform analysis on a data set that is being held by a health care facility. The data contains very sensitive medical information about the facility's patients. The facility cannot provide the real data to the research
>
> her but is willing to take the researchers' SAS program, run it on their data and send him/her the results back. The only problem is that the researcher cannot write a SAS program that s/he knows will work on the facility's data set without first knowing what exactly the data looks like.

There are at least three ways to satisfy these kinds of requirements:

1. Remove all of the identifying information from the data.
2. Encrypt all of the identifying information in the data, which is functionally equivalent to the first option.
3. Anonymize the identifying information through randomization.

The first two options will not actually meet the needs of our two scenarios. In the first scenario the test team needs realistic data which

CHEO Research Institute © Copyright 2007. All Rights Reserved

includes names and addresses, otherwise they would not be able to test the data processing application properly. For example, the application may not work properly with names having special characters (e.g., French names with accents). The only way testing will discover that bug is if a large number of realistic customer names are tried and some of these are names with special characters. So removing the names or encrypting them will not do. In the second scenario the researcher needs to get realistic data to write his/her SAS program. If the names are removed s/he may write a program that works with data sets without names and addresses, but then the program may not actually work with the real data.

Anonymization through randomization is the functionality which we will be demonstrating in this case study. The basic idea is that the PrivacyAnalytics systems will replace the real names and addresses with bogus names and addresses. The bogus names and addresses are taken from a large database of real Canadian names and addresses. The system ensures that all of that information looks real (for example, if it replaces a male name with a randomly selected name from its database, it will also be a male name).

CHEO Research Institute © Copyright 2007. All Rights Reserved

Randomizing Names and Addresses

In this case study, a 38,296-record dataset comprised of members of the College of Physicians and Surgeons of Ontario is subjected to name and postal code randomization in PrivacyAnalytics.

As a first step, the dataset is converted to comma delimited format and is opened in PrivacyAnalytics.

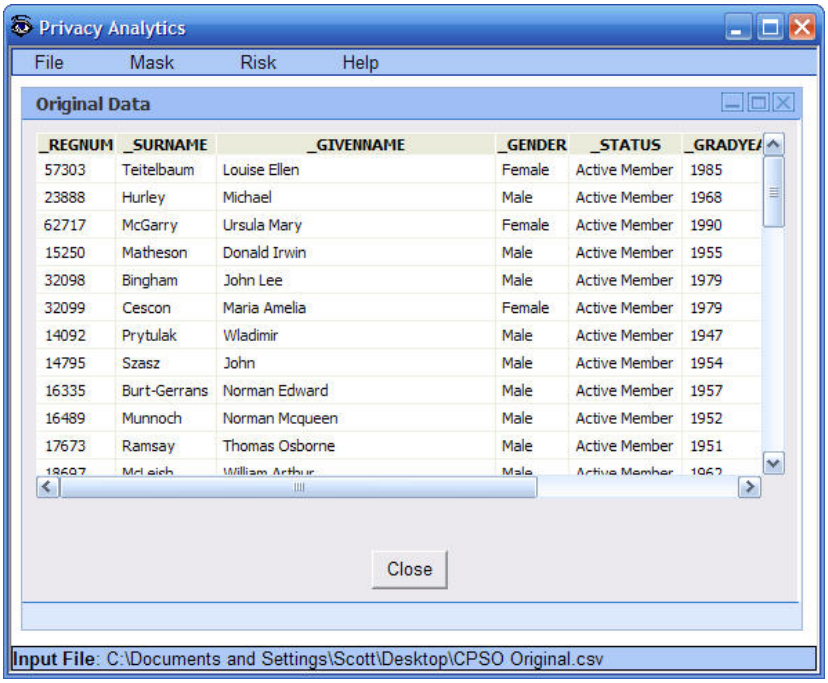


Figure 1: Original data from the College of Physicians and Surgeons of Ontario members' database.

The *Randomize Names* dialogue box (Figure 2) appears when the *Randomize Names* function is selected from the *Mask* menu. As name randomization is the first operation to be performed on the dataset, we leave the radio button beside *Original Input File* selected.

When randomizing names, we initially select the dataset's first name variable. This tells PrivacyAnalytics which variable in the dataset corresponds to the record's first name and should therefore be altered

CHEO Research Institute © Copyright 2007. All Rights Reserved

during the operation. Secondly, we select the gender variable. This tells PrivacyAnalytics which variable corresponds to the gender of the record. Finally, we select the value in the gender variable which corresponds to each gender. PrivacyAnalytics can produce male and female first names. A typical gender variable might have “male”, “female” and “unknown” values.

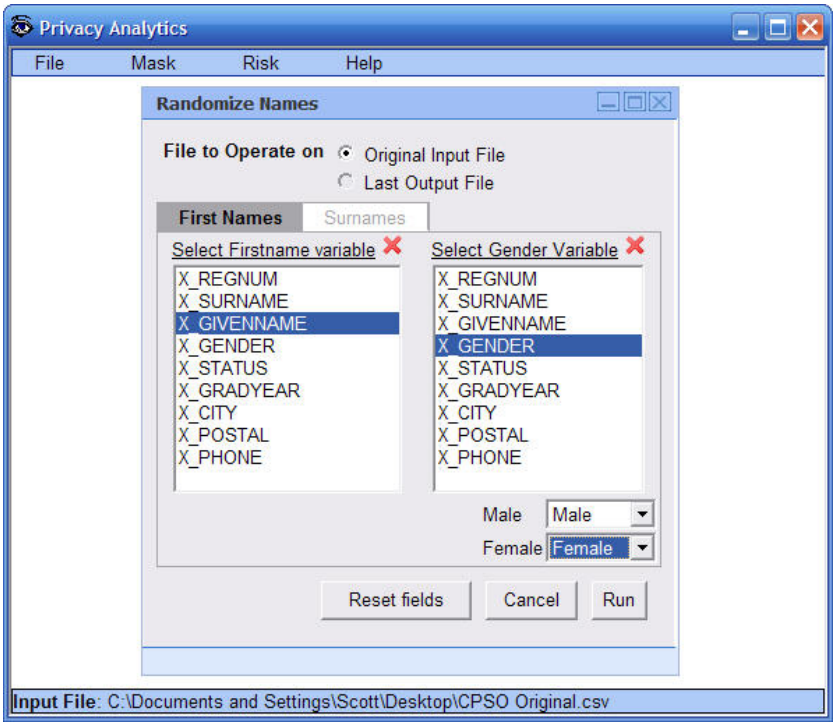


Figure 2: Randomize names (First Names Tab).

By selecting the *Surnames* tab (Figure 3), we are able to further mask the dataset by randomizing surnames. Unlike first name randomization, surnames are not gender-based.

We select the dataset’s surname variable. This tells PrivacyAnalytics which variable in the dataset corresponds to the record’s surname and should therefore be altered during the operation. If left blank, PrivacyAnalytics would conduct only first name randomization.

CHEO Research Institute © Copyright 2007. All Rights Reserved

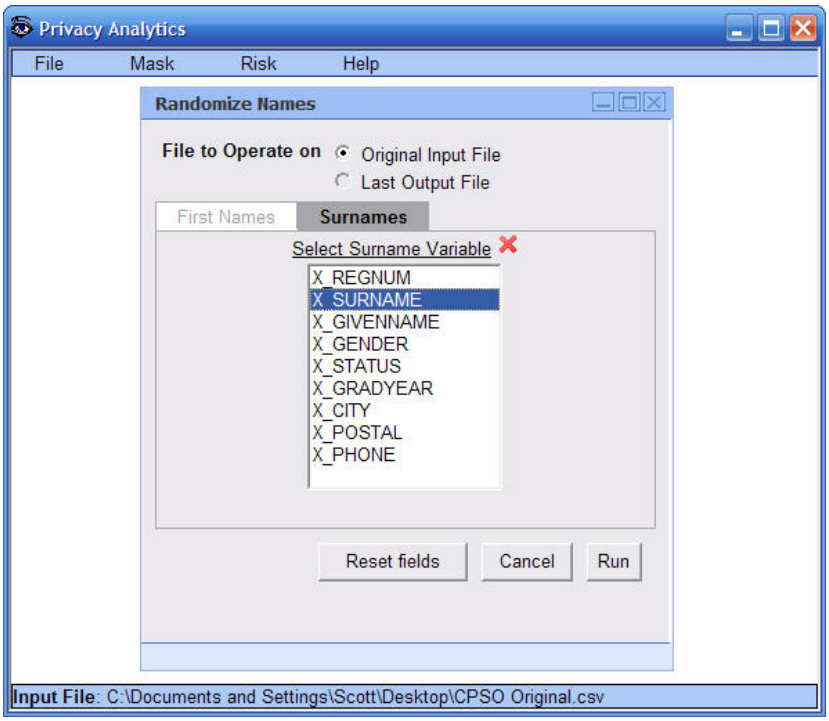


Figure 3: Randomize Names (Surnames Tab).

When these options have been selected, we press *Run* to conduct the randomization. A progress bar (Figure 4) will appear to indicate PrivacyAnalytic's progress.

CHEO Research Institute © Copyright 2007. All Rights Reserved

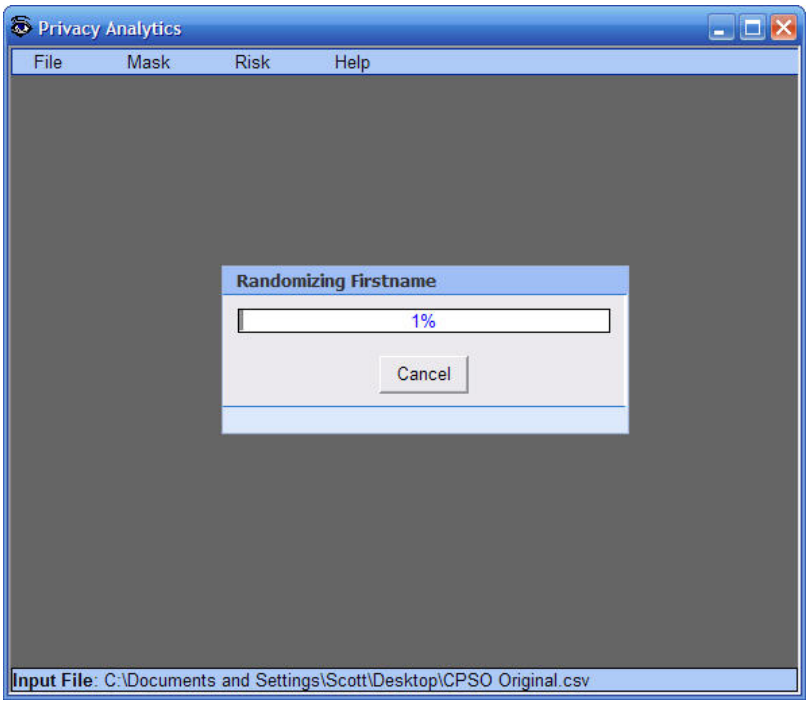
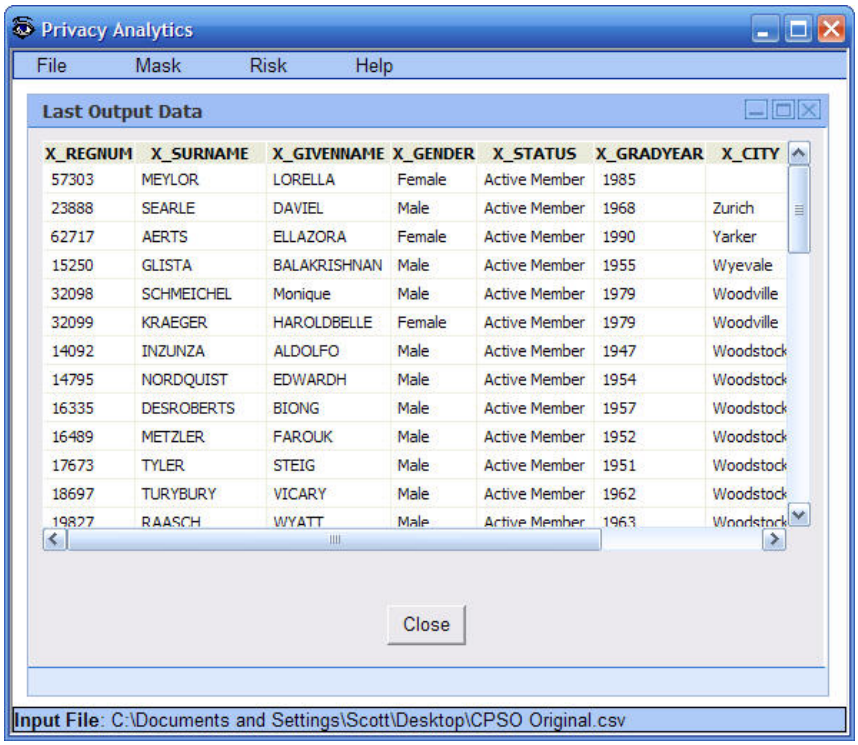


Figure 4: Progress bar.

When first name and surname randomization is complete, a notification dialogue box will appear. By selecting the option to *View Last Output Data* from the *File* menu, you can see the changes to the dataset (Figure 5).

Note in Figure 5 that all surnames have been randomized, while first names have been randomized only for those values where a “male” or “female” gender value could be assigned. Therefore, the records with the “unknown” value for gender will not be affected.

CHEO Research Institute © Copyright 2007. All Rights Reserved

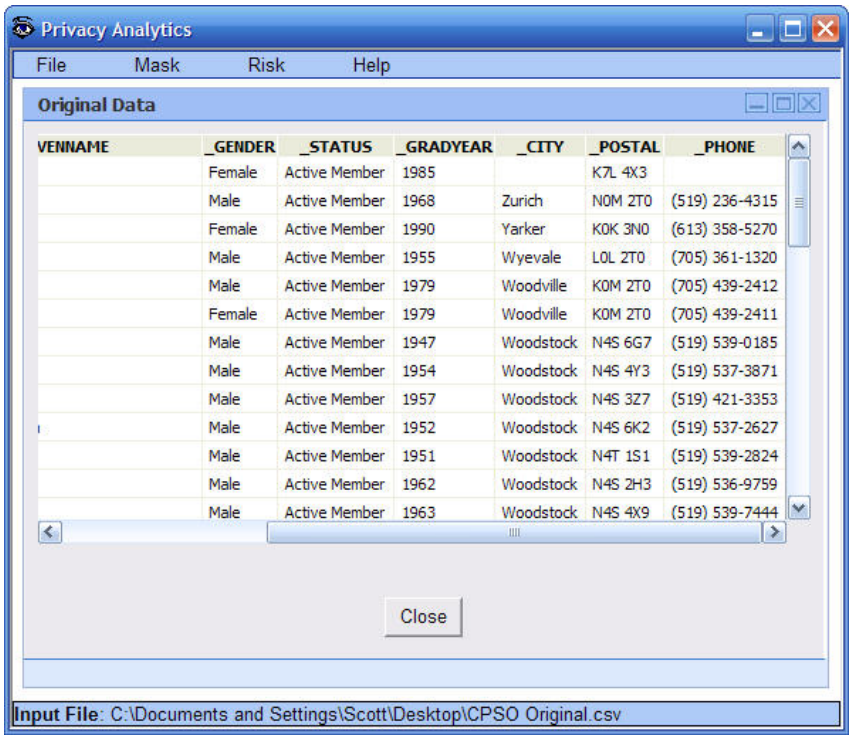


Figures 5: Last Output Data.

The next step in this case study is to randomize the postal codes associated with each record. We will show a basic type of postal code randomization in this case study.

A screen shot of the postal codes for our data set before randomization is shown in Figure 6. To initiate postal code randomization, we select *Randomize Postal Codes* from the *Mask* menu. The associated dialogue box appears (as shown in Figure 7).

CHEO Research Institute © Copyright 2007. All Rights Reserved



Figures 6: The postal code information before randomization.

Since our name randomization produced last output data, we need to change the radio button to indicate that we will perform the operation on this data. We then select the dataset's postal code variable, which will be randomized in this operation. Finally, we select whether we would like to randomize by Local Delivery Unit (LDU), Province, or All and press *Run* to start the operation. In this case we will select All. Selecting All means that the full postal code will be replaced by some other valid postal code in Canada. No attempt will be made to preserve the same Province or the Forward Sortation Area (the first three characters of the postal code).

CHEO Research Institute © Copyright 2007. All Rights Reserved

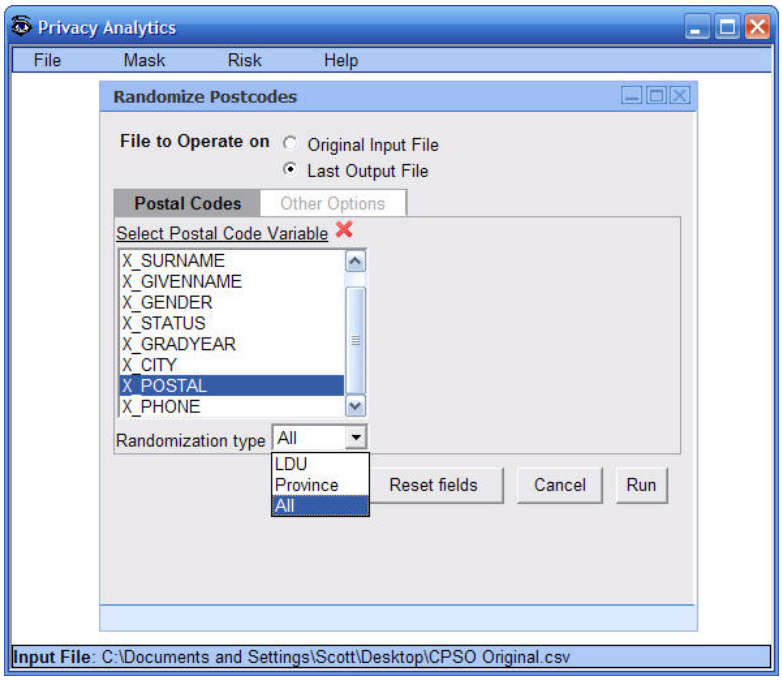


Figure 7: Randomize Postal Codes.

As in the first step, a progress bar will appear. When the operation is complete, the last output data can similarly be viewed (Figure 8), and the new randomized data can be saved (Figures 9 and 10).

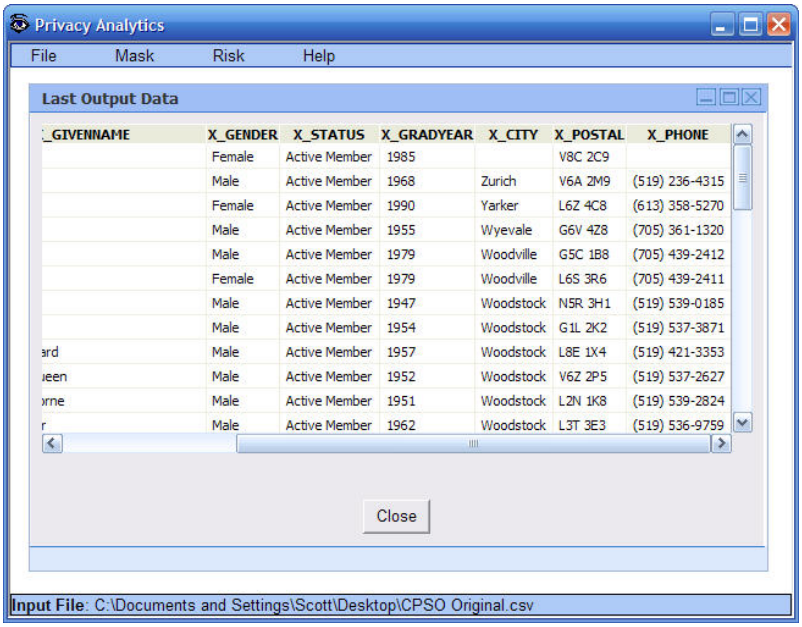


Figure 8: The randomized postal codes.

CHEO Research Institute © Copyright 2007. All Rights Reserved

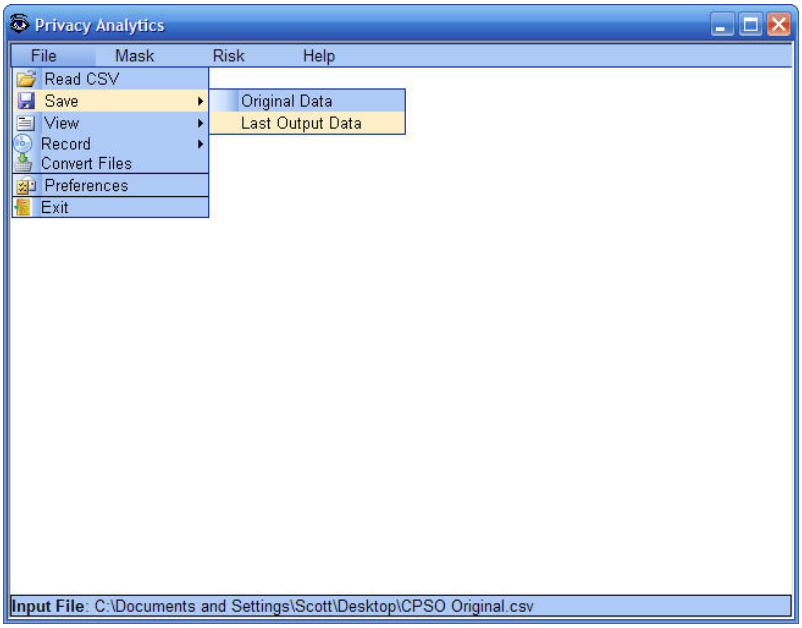


Figure 9: Save Last Output Data.

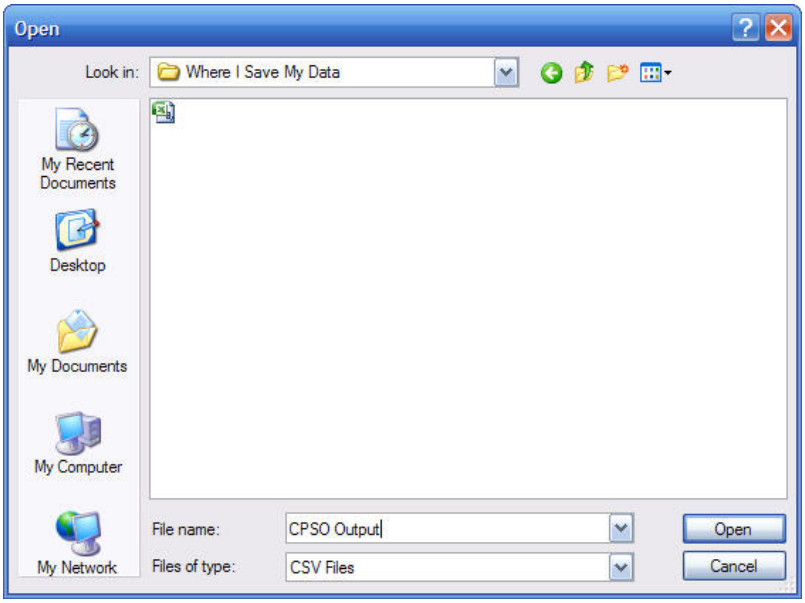


Figure 10: Save dialogue box.

CHEO Research Institute © Copyright 2007. All Rights Reserved