

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

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).

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.

File	Mask	Risk Help			
Original C)ata				
REGNUM	_SURNAME	_GIVENNAME	_GENDER	_STATUS	_GRADYE/
57303	Teitelbaum	Louise Ellen	Female	Active Member	1985
23888	Hurley	Michael	Male	Active Member	1968
62717	McGarry	Ursula Mary	Female	Active Member	1990
15250	Matheson	Donald Irwin	Male	Active Member	1955
32098	Bingham	John Lee	Male	Active Member	1979
32099	Cescon	Maria Amelia	Female	Active Member	1979
14092	Prytulak	Wladimir	Male	Active Member	1947
14795	Szasz	John	Male	Active Member	1954
16335	Burt-Gerrans	Norman Edward	Male	Active Member	1957
16489	Munnoch	Norman Mcgueen	Male	Active Member	1952
17673	Ramsay	Thomas Osborne	Male	Active Member	1951
18607	Mrt eich	William Arthur	Mala	Active Member	1067
•					
		CI	ose		

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

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.

e	Mask Risk Help	
	Randomize Names	
	File to Operate on <a>O O	iginal Input File st Output File
	First Names Surnan	es
	Select Firstname variable X_REGNUM X_SURNAME X_GIVENNAME X_GENDER X_STATUS X_GRADYEAR X_CITY X_POSTAL X_PHONE	Select Gender Variable X X_REGNUM X_SURNAME X_GIVENNAME X_GENDER X_GRADYEAR X_CITY X_POSTAL X_PHONE
		Male Male ▼ Female Female ▼
	Res	et fields Cancel Run

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.

ïle	Mask Risk	Help	
	Randomize Nar	nes	
	File to Operat	e on ⓒ Original Input Fil C Last Output File	e
	First Names	Surnames	
		Select Surname Variable	
		Reset fields	Cancel Run

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.

Se Privacy Analytics				_ 🗆 🔀
File Mask	Risk	Help		
	Randomi	izing Firstname		
		1%		
		Cancel		
Input File: C:\Documen	ts and Setting	s\Scott\Desktop\CPS0	Original.csv	

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.

lie	IVIASK P	usk neij	,				
Last Outp	ut Data						X
X_REGNUM	X_SURNAME	X_GIVENNAME	X_GENDER	X_STATUS	X_GRADYEAR	X_CITY	~
57303	MEYLOR	LORELLA	Female	Active Member	1985		
23888	SEARLE	DAVIEL	Male	Active Member	1968	Zurich	
62717	AERTS	ELLAZORA	Female	Active Member	1990	Yarker	
15250	GLISTA	BALAKRISHNAN	Male	Active Member	1955	Wyevale	
32098	SCHMEICHEL	Monique	Male	Active Member	1979	Woodville	
32099	KRAEGER	HAROLDBELLE	Female	Active Member	1979	Woodville	
14092	INZUNZA	ALDOLFO	Male	Active Member	1947	Woodstock	
14795	NORDQUIST	EDWARDH	Male	Active Member	1954	Woodstock	
16335	DESROBERTS	BIONG	Male	Active Member	1957	Woodstock	
16489	METZLER	FAROUK	Male	Active Member	1952	Woodstock	
17673	TYLER	STEIG	Male	Active Member	1951	Woodstock	
18697	TURYBURY	VICARY	Male	Active Member	1962	Woodstock	_
19827 <	RAASCH	WYATT	Male	Active Member	1963	Woodstock	~
			Close				

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).

	-ividon.	TMO	x neip					
Original I	Data							
VENNAME		_GENDER	_STATUS	_GRADYEAR	_CITY	_POSTAL	_PHONE	
		Female	Active Member	1985		K7L 4X3		
		Male	Active Member	1968	Zurich	NOM 2TO	(519) 236-4315	=
		Female	Active Member	1990	Yarker	KOK 3NO	(613) 358-5270	
		Male	Active Member	1955	Wyevale	LOL 2TO	(705) 361-1320	-
		Male	Active Member	1979	Woodville	KOM 2TO	(705) 439-2412	
		Female	Active Member	1979	Woodville	KOM 2TO	(705) 439-2411	
		Male	Active Member	1947	Woodstock	N4S 6G7	(519) 539-0185	
		Male	Active Member	1954	Woodstock	N4S 4Y3	(519) 537-3871	
		Male	Active Member	1957	Woodstock	N4S 3Z7	(519) 421-3353	
		Male	Active Member	1952	Woodstock	N4S 6K2	(519) 537-2627	
		Male	Active Member	1951	Woodstock	N4T 1S1	(519) 539-2824	
		Male	Active Member	1962	Woodstock	N4S 2H3	(519) 536-9759	
		Male	Active Member	1963	Woodstock	N4S 4X9	(519) 539-7444	~
<					JIII		>	

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).

ile	Mask Risk	: Help			
	Randomize Posto	odes			
	File to Operate o	n Original C Driginal	Input File put File		
	Postal Codes	Other Option	ıs		
	Select Postal Code	Variable X			
	X_SURNAME X_GIVENNAME X_GENDER X_STATUS X_GRADYEAR X_CITY X_POSTAL X_PHONE				
	Randomization type	e All 🗾			
		LDU Province All	Reset fields	Cancel	Run
	-				

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).

ile	Mask	Risk	Help					
Last 0	utput Data])
GIVEN	INAME	X_GENDER	X_STATUS	X_GRADYEAR	X_CITY	X_POSTAL	X_PHONE	
		Female	Active Member	1985		V8C 2C9		
		Male	Active Member	1968	Zurich	V6A 2M9	(519) 236-4315	III
		Female	Active Member	1990	Yarker	L6Z 4C8	(613) 358-5270	
		Male	Active Member	1955	Wyevale	G6V 4Z8	(705) 361-1320	
		Male	Active Member	1979	Woodville	G5C 1B8	(705) 439-2412	
		Female	Active Member	1979	Woodville	L6S 3R6	(705) 439-2411	
		Male	Active Member	1947	Woodstock	N5R 3H1	(519) 539-0185	
		Male	Active Member	1954	Woodstock	G1L 2K2	(519) 537-3871	
ard		Male	Active Member	1957	Woodstock	L8E 1X4	(519) 421-3353	
Jeen		Male	Active Member	1952	Woodstock	V6Z 2P5	(519) 537-2627	
me		Male	Active Member	1951	Woodstock	L2N 1K8	(519) 539-2824	
r		Male	Active Member	1962	Woodstock	L3T 3E3	(519) 536-9759	~
<					1		>	
			L	Close				

Figure 8: The randomized postal codes.



Figure 9: Save Last Output Data.



Figure 10: Save dialogue box.