Anonymizing Spreadsheet Data and Metadata with AnonymousXL

Joeri van Veen
Infotron
Delft, the Netherlands
joeri@infotron.nl

Felienne Hermans
Delft University of Technology
Delft, the Netherlands
f.f.j.hermans@tudelft.nl

ABSTRACT

In spreadsheet risk analysis, we often encounter spreadsheets that are confidential. This might hinder adoption of spreadsheet analysis tools, especially web-based ones, as users do not want to have their confidential spreadsheets analyzed. To address this problem, we have developed AnonymousXL, an Excel plugin that makes spreadsheets anonymous with two actions: 1) remove all sensitive metadata and 2) obfuscate all spreadsheet data within the Excel worksheets such that it resembles, untraceably, the original values.

1. INTRODUCTION

When commercializing our Breviz analysis toolkit [2, 3, 4] as an online tool called PerfectXL, we ran into the problem that customers often do not want to upload, share or even show us confidential spreadsheets. Therefore, we have developed a tool that obfuscates [1] both the data and the metadata in a spreadsheet, while the values still resemble the original ones. By construction, we guarantee that our anonymization does not create or resolve Excel errors. This enables us to run our smell detection tool on the anonymized spreadsheets as if we were analyzing the original. This paper describes the capabilities, limitations and applications of AnonymousXL.

2. METADATA REMOVAL

AnonymousXL removes spreadsheet metadata: the author, the date the file was last opened and the total edit time, in order to remove any ties with the company that the spreadsheet originally came from. In addition, worksheet names within the spreadsheet are replaced with anonymous names.

2.1 Numerical and Date Related Metadata

All numerical metadata information is converted to 0. At this time, only the numerical metadata "revision number" and "total editing time" are converted. All metadata that is of a date type is set to the day of anonymization: "last print date", "creation date" and "last save time".

2.2 Textual Metadata

The following textual metadata are set to the text string "anonymous": title, subject, author, keywords, comments, template, last author, application name, security, category, format, manager, company.

3. DATA OBFUSCATION

Data obfuscation is the alteration of data to make it anonymous. This happens linearly, from the first sheet to the last sheet, from the first to the last cell of the used range of cells in each worksheet. We use different techniques for different types of data in the spreadsheet: numeric data, dates, textual data, formulas and other types of data.

3.1 Numeric Data

The basic step for anonymizing a number is to randomly add or subtract up to 60% of its original value. Or, mathematically, for any number N in a cell, N is replaced by $N \pm N \times 0.6 \times r$ where r is a random value in the range [0, 1]. We treat integers and real numbers differently: Integer values remain integer, real numbers keep their decimals.

There is one exception in the anonymization: In PerfectXL, one of the analyses that is performed is the occurrence of so-called 'magic numbers', numbers of which the meaning might be unclear to the user. There are some numbers, however, that are not considered to be magical, because of their frequent occurrence: 0, 1, 2, 12, 100, 365, 1000. Therefore, these numbers remain as is in our anonymization process. Since all text fields (including column names) get changed, we believe that leaving the non-magic numbers intact does not pose a threat to the anonymity of the spreadsheet, since labels give numbers semantics.

3.2 Dates

Dates are converted into random dates in the range of representable dates in VBA, in contrast with metadata, in which all date values are set to the day of anonymization. This randomness is introduced as to maintain data variation.

3.3 Textual Data

For textual data, it does not suffice to simply change all textual values to "text", since in many situations, it matters to keep equal strings equal. An example of such a situation is a pivot table, as shown in Figure 1. Should we change all categories to "text", the spreadsheet would not work any-

more, as pivot tables cannot contain two fields of the same name. If we would replace all textual values by unique ones, such as "text1", "text2", "text3", as shown in Figure 2, it does work, pivot tables however are often based on textual data (which denote categories, for example). This means that where there once were three categories ("a", "b" and "c", in Figure 1), now there are many (eight different ones in Figure 2). Pivot tables calculate their size based on the number of unique values they find for a category, so pivot tables become larger than they were originally. This can lead to problems, since multiple pivot tables are often situated close to each other on the same worksheet. If the pivot tables grow because of the anonymization, they can start to overlap and unfortunately, this causes Excel to crash.

	A	B	C	D	E
1	Category	Count		Row Labels	Sum of Count
2	b	20		a	53
3	b	41		b	144
4	a	29		c	179
5	b	53		Grand Total	376
6	c	88			
7	a	24			
8	c	91			
9	b	30			

Figure 1: Original spreadsheet with three categories

	A	B	C	D	E
1	Category	Count		Row Labels	Sum of Count
2	text1	20		text1	20
3	text2	41		text2	41
4	text3	29		text3	29
5	text4	53		text4	53
6	text5	88		text5	88
7	text6	24		text6	24
8	text7	91		text7	91
9	text8	30		text8	30
10				Grand Total	376

Figure 2: Simple text replacement

	A	B	C	D	E
1	unique 1	unique 2		unique 3	Sum of unique 2
2	unique 5	28		unique 5	102
3	unique 5	48		unique 6	52
4	unique 6	18		unique 7	151
5	unique 5	1		unique 8	305
6	unique 7	143			
7	unique 6	34			
8	unique 7	8			
9	unique 5	25			

Figure 3: AnonymousXL applied to table

Therefore, we anonymize all textual values while keeping intact cell uniqueness by replacing texts with "unique1", "unique2", "unique3", etc. (for example, "unique6" represents the textual value "a" in Figure 3).

3.4 Formulas

Formulas are basially left alone. The only modification made to formulas are sheet references, since sheet names are made anonymous as well.

3.5 Other Types

Other data types usually fall under either categories mentioned (for instance, a currency type is simply considered a number). A special note on booleans TRUE and FALSE: as booleans are interpreted by Excel as 0 and 1, they are not changed. However, booleans are seldom present as literal values. They are often the result of formulas, in which case they only change in accordance with modifications to the data they depend on.

4. INTRODUCING EXCEL ERRORS

By changing data in Excel cells, errors might be induced that were not present in the original spreadsheet. For instance, in the formula =A1/(3-A2), division by zero might occur (and thus be reported after analysis) if A2 becomes 3, which could happen because of the anonymization step in which data in cells is decreased or increased by 60% of their original value.

To resolve this, we save the list of all formulas that result in an error before the anonymization. Then, after we anonymize each data cell, we verify that we have not changed this list. For this, we do not have to analyze all formulas in the spreadsheet, we only analyze the recursive precedents of the cell, plus all formulas that contain the INDIRECT function.

5. LIMITATIONS

5.1 Confidential formulas

Every so often, spreadsheets contain confidential formulas. All formulas, including those confidential ones, are left unaltered to preserve analysis results. This might not be sufficient for some users.

5.2 Embedded constants

In the current implementation, we only change numeric values in cells and not within formulas, such as in =SUM(A1:A10)*1.2. This is a limitation because these constants too can be of importance to the spreadsheet owner and thus confidential.

5.3 Analysis Types

Different kinds of spreadsheet analyses scan for different kinds of patterns. Developed to complement PerfectXL, AnonymousXL leaves intact formulas, boolean literals and certain numbers for they are key to mimicking analysis of the original spreadsheet. Nevertheless, AnonymousXL or a slight variation of it could carry great potential for alternative analysis types.

6. REFERENCES

[1] D. E. Bakken, R. Parameswaran, D. M. Blough, A. A. Franz, and T. J. Palmer. Data obfuscation: Anonymity and desensitization of usable data sets. *IEEE Security & Privacy*, 2(6):34–41, 2004.

[2] F. Hermans, M. Pinzger, and A. van Deursen. Supporting professional spreadsheet users by generating leveled dataflow diagrams. In *Proc. of ICSE '11*, pages 451–460, 2011.

[3] F. Hermans, M. Pinzger, and A. van Deursen. Detecting and visualizing inter-worksheet smells in spreadsheets. In *Proc of ICSE '12*, pages 441–451, 2012.

[4] F. Hermans, M. Pinzger, and A. van Deursen. Detecting code smells in spreadsheet formulas. In *Proc of ICSM '12*, pages 409–418, 2012.