VBA in the spreadsheets from the FUSE corpus

Patrick O'Beirne, Systems Modelling Ltd
Pob-at-sysmod-.-com

FUSE spreadsheets corpus

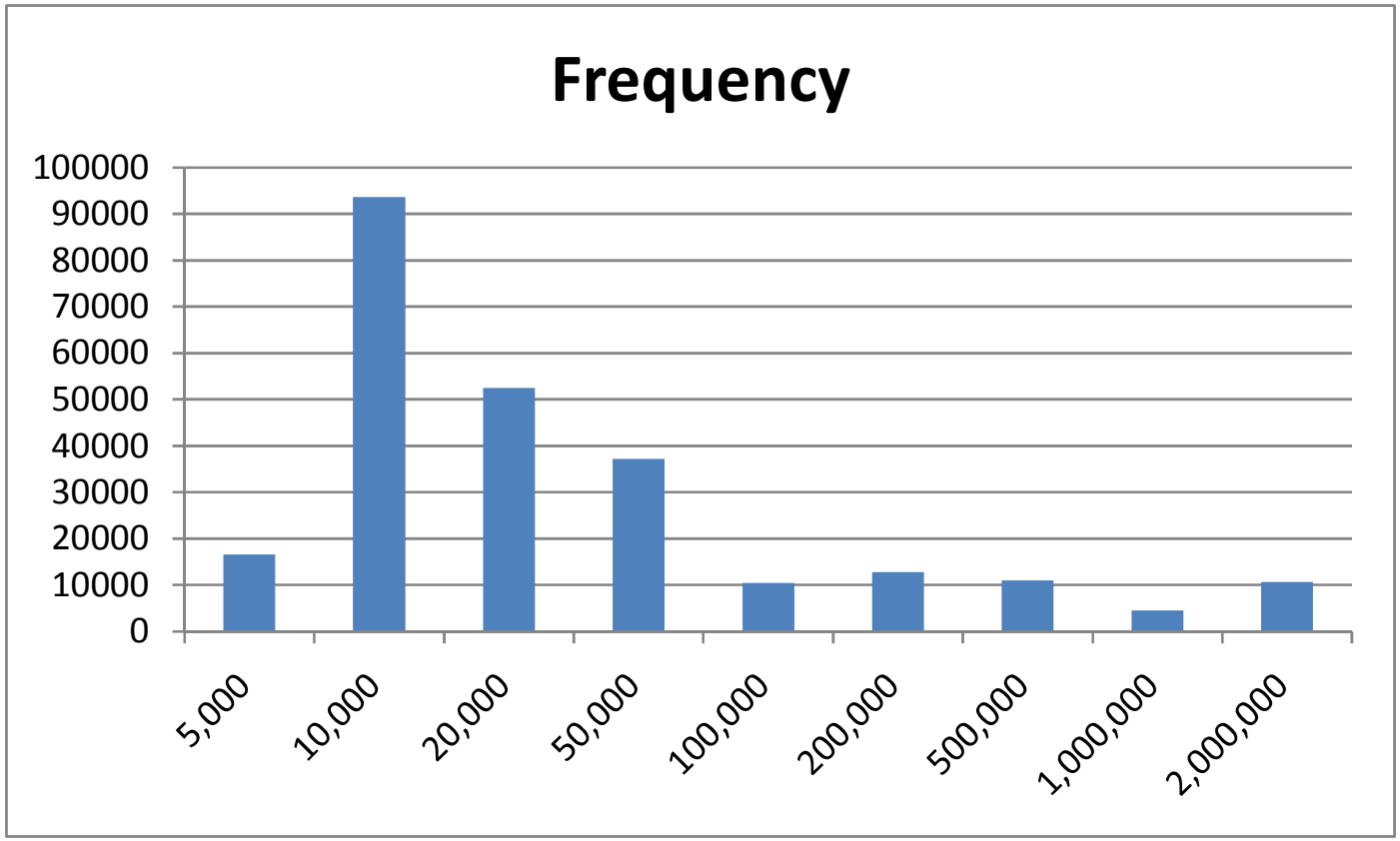
- Fuse: A Reproducible, Extendable, Internet-scale Corpus of Spreadsheets
- static.barik.net/barik/publications/msr2015/PID3640389.pdf
- Downloads: http://static.barik.net/fuse/
- Their analysis using Java POI:
- fuse-bin.analysis.dedup.poi-dec2014.json.gz (298MB)
- Archive of 249,376 deduped binary spreadsheets:
- fuse-binaries-dec2014.tar.gz (6.9 GB)

Basic statistics

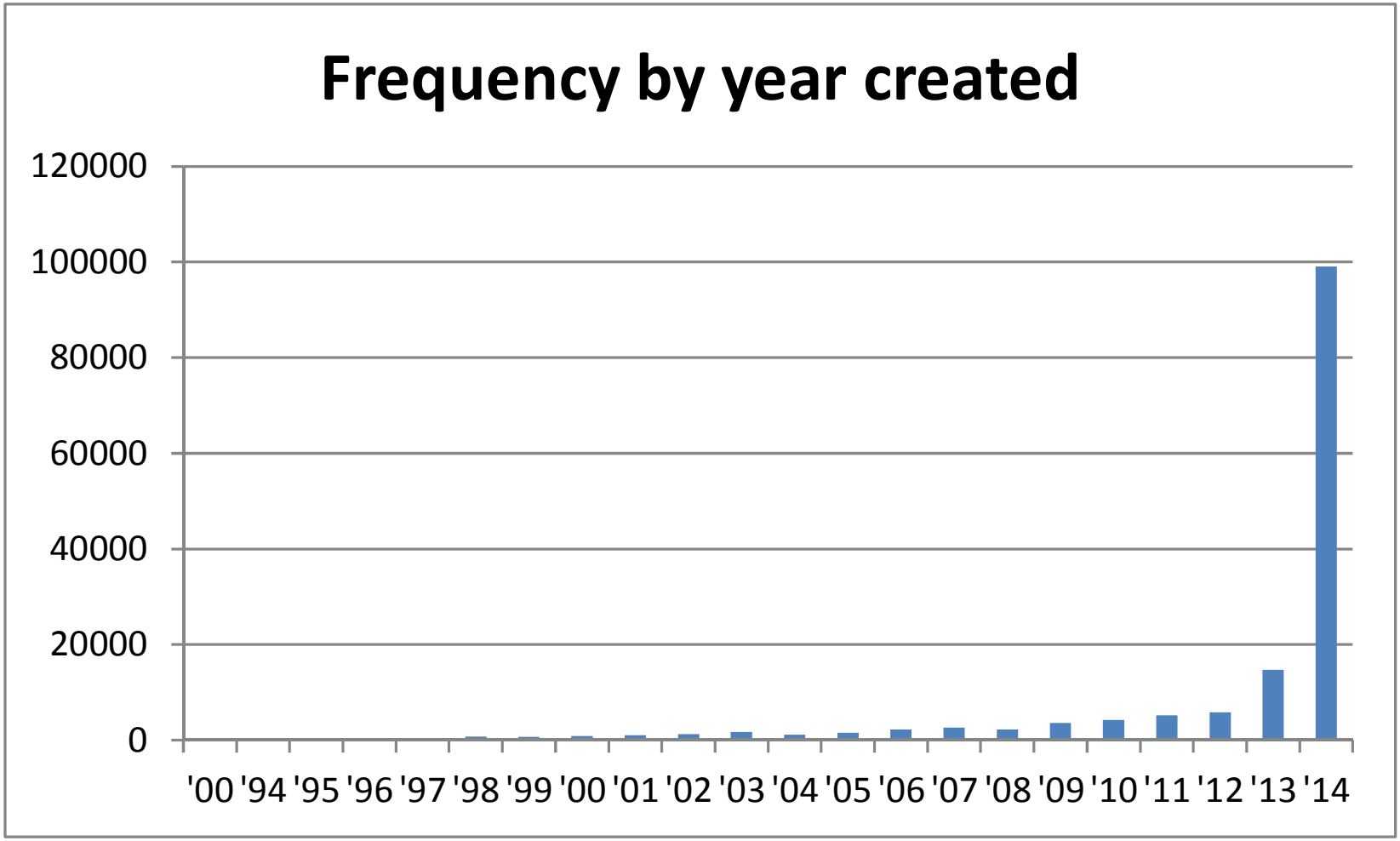
- 249,376 files, hashed names, no extensions, in one folder
- All have unique SHA1 hashes.
- 238,665 have DocFile header (Excel pre 2007)
  - POI analysis assigns filetype .xls
  - 362 were Template files, so extension should have been .xlt;
  - 5 were Addin files, so extension should have been .xla
- 10,703 are Zip files (Excel 2007+)
  - They give them filetype .xlsx, no .xlsm found
- 7 are invalid
- Use DSOfile.dll to get properties (DateCreated etc)
- 553 had no BIFF header; 346 are Google docs/spreadsheets

249,376 Excel files, 3K to 1MB

- 1MB limit is cutoff point for Common Crawler database
- I have not downloaded originals to check actual max size



Most are recent



Top 20 functions in formulas

Function	Sum all usage	Function	Files using
IF	2,496,350	SUM	8,760
SUM	1,225,228	IF	3,174
ISBLANK	807,679	HYPERLINK	1,655
AVERAGE	446,179	AVERAGE	1,116
VLOOKUP	430,388	VLOOKUP	1,003
ISTEXT	329,433	ISBLANK	772
ROUND	204,625	MAX	757
LEFT	187,699	ROUND	721
IFERROR	181,976	MIN	570
VALUE	166,095	AND	545
SUBSTITUTE	162,372	IF	501
OR	151,397	OR	392
HLOOKUP	132,472	INDEX	367
INDEX	113,415	COUNT	356
SUMIF	94,978	LEFT	311
AND	91,634	MATCH	307
NOT	77,261	CONCATENATE	305
MAX	67,775	VALUE	242
HYPERLINK	66,157	ISTEXT	194
SQRT	60,802	{ARRAY}	187

Miscellaneous observations

- Excel links: 1055(0.4%) had links; 772 had 1 link, the maximum was 166 links.
- Defined names: 24569 (10%) had names, 8757 had just one name, the most found was 7843 names (of which 7842 were constants).
- Author name: 136127 non-blank names (55%); 106272 were "International Triathlon Union", the major source.
- 78% had one used worksheet; the highest was 147.
- 234593 (94%) had no formulas at all. The highest one had 58389 formula cells.

Cautions with interpretation

- Just because a workbook contains Excel #error values does not mean that it contains errors in the sense of defects (bugs) caused by mistakes.
  - One workbook contains 9,100 error values, but these include 6 sheets of #N/A values from VLOOKUP of blanks. These values are expected and therefore the user would not regard these as *errors*.
- I would look more closely at #REF error values

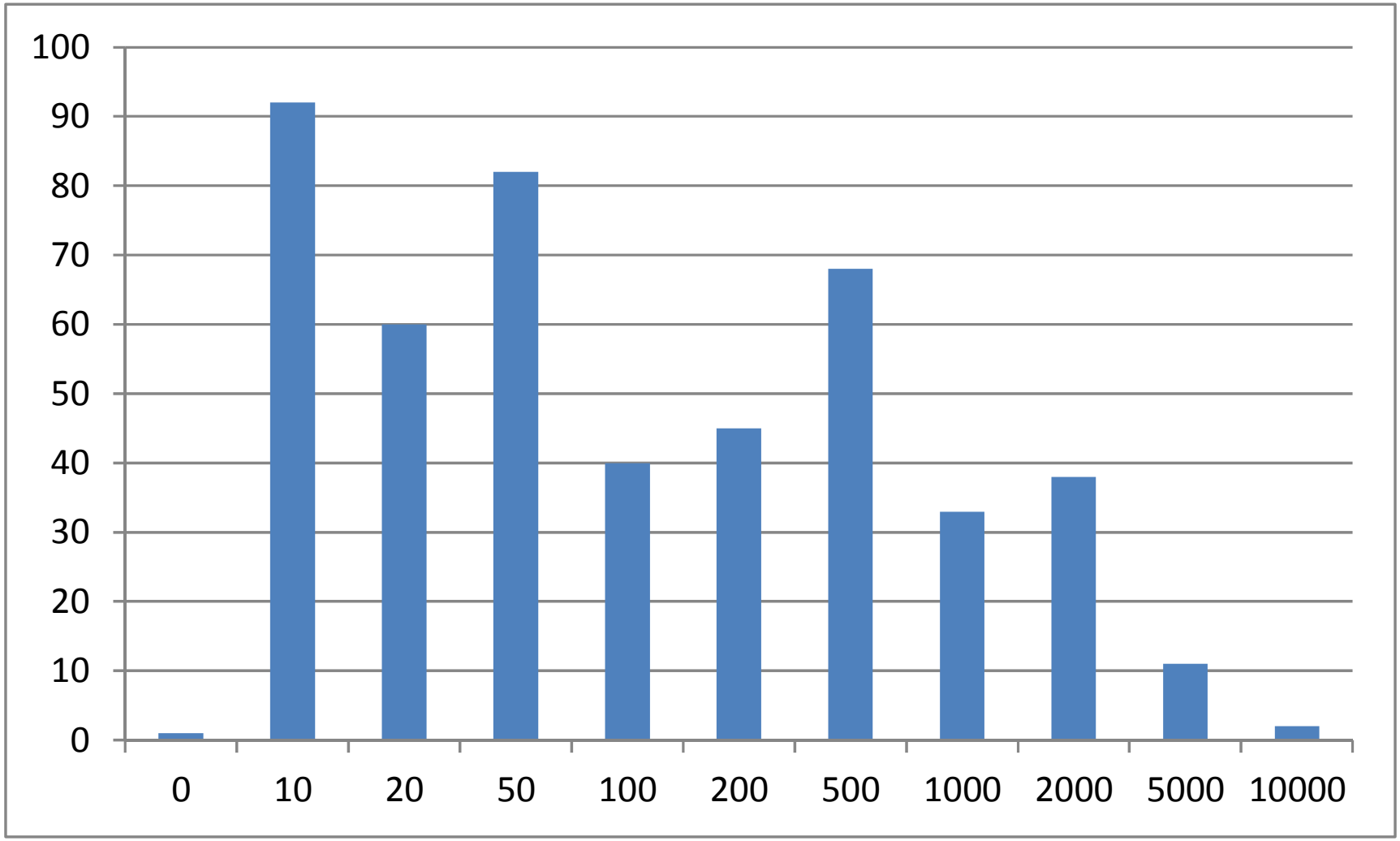
Extracting the VBA

- Use decalage.info Oletools Olevba.py to extract VBA stream
- 5,358 contain a VBA component
- 3,258 could not be analysed because corrupt
- MD5 hash shows 472 unique VBA code files out of 2,100

Top 10 sites represent 22% of the 472 files.

Site	Files
birmingham.gov.uk	22
schools.utah.gov	17
portal.ncdenr.org	11
transportation.ky.gov	8
portal.hud.gov	8
epublications.bond.edu.au	8
fina.hr	8
mahoningcountyoh.gov	8
ohr.edu	7
liferaydemo.unl.edu	6

Lines of VBA Code per workbook



Macro recording or programming?

- Of the 427 unique VBA files:
- How much VBA is macro recorded?
  - 102 have "Macro recorded by…" and no Dim statements
  - Frequently associated with use of .Select
- What level of skill is shown?
  - All – from simple recorded macros to user-defined classes and WinAPI calls

Indicators of code quality in the corpus

- Only 78 of 472 have Option Explicit
- 7 use Option Base
- 6 use Option Private
- Only 1 uses DoEvents
- 50 use On Error GoTo
- 59 use On Error Resume Next
- 8 use ByRef

Top 10 common words in code

(excluding common VBA reserved words)

Variable	Count of files
i	162
CommandButton1_Click	128
Target	105
Macro1	72
x	68
Shapes	59
iMax	57
xlLinkTypeExcelLinks	53
newPath	53
alinks	53

Variable	Sum of Usage
i	4305
parametri	1804
nLinea	1385
Sheet2	1304
j	1060
Sheet3	886
Counter	842
DestSht	825
oRango	810
NewSht	803

"I" is the variable most commonly used in VBA.

Usage count is skewed by repetition

- eg "parametri" is an array name used 1804 times in 2 files

Further research

- What would be interesting?

  - Patrick O'Beirne, Systems Modelling Ltd
  - www.sysmod.com
  - pob@ that domain
  - sysmod.wordpress.com
  - @ExcelAnalytics