Exemplar -- Code Comparison

Cosgrove Computer Systems Inc.
7411 Earldom Avenue
Playa del Rey, California 90293-8058
(310) 823-9448
JCosgrove@Computer.org, www.CosgroveComputer.com

5/16/2003

Esq.
Law Offices

Ref: XYZ v ABC Code comparison

Dear Mr. Esq:

Attached is the initial draft of the report on the comparison of the two versions of the program source code which compiles the data from physical fitness of students. These are respectively titled "M" for Dr. M's version and "L" for the original version submitted by that office. Additional details from the analysis are contained in the accompanying Excel spread-sheet (SS) which supported the analysis and was used to compile the statistics quoted in the conclusions.

Summary:
The M version of the software appears to be effectively a complete derivative of the L version for the following reasons:

1. Once the cosmetic (see below) changes are eliminated, there is an effective 98% match between the L and the seemingly derived M version.
2. Two obvious errors exist in both versions which – though trivial in their operational impact – effectively show that a copy was made as opposed to a new development.
3. The nature of the cosmetic changes implies a conscious attempt to disguise the similarity of the two programs for some unknown purpose.

Disputed Issues
In the M complaint, it is stated that "Dr. M totally modified and re-wrote the SAS language computer software that he had initially created."

The L position is that the M version is a derivative of the L version.

The results of our investigation suggest very strongly that the L version was the source of the M version, and that while a number of cosmetic changes were made; the claim that the software was totally modified and rewritten is not supported by the contents of the files submitted.

Replicated Errors
Two obvious, minor errors were present in both versions. In some professions (e.g., map making), it is common practice to embed a small error in the work so that copyright

infringement can be shown if the same error also appears in an allegedly original work. In this case, although it is unlikely that the errors were intentional, they serve the same purpose – effectively proving that a copy was made.

- It is highly unlikely that one could do a total modification or rewrite and have the conversion from inches to meters as 39.47 (39.**3**7 is the correct number) in the Body Mass Index calculation. This error appears in both the L version and the M file.
- Also, after the Y2K publicity, it is also highly unlikely that one would calculate YEAR4=YEAR2+1900 rather than changing the input to a 4-digit field. While that is curious, what is more interesting is why the M file would define a variable YEAR4 at the precise point in the file as L does, and **then never use it**. This error does not support independence in the M development.

Cosmetic Changes

The nature of source code editors commonly used in software development makes certain kinds of changes simple to perform. The relevant types of changes indicated in the M version of the software can be accomplished quite easily and most likely were accomplished in less than a day's effort. In fact, as a trial, all of the group changes of names were done by our analyst in less than 15 minutes. It may be helpful to think of the significance of name changes in a copyrighted work as being the same as taking a literary work and changing the names and locales of the main characters but keeping the plot line and most of the character interaction. Nobody could argue that one is absolved of violating a copyright by such a cosmetic change.

The types of cosmetic changes identified in our analysis are as follows:

1. Group changes of names – For example "sex" was changed to "gender". Twenty variable names were thus changed.
2. Re-ordering of code sections which are not dependent on the order of the statements – This was done in the 1st ~100 lines of the software for some reason. One likely purpose is to create an initial impression on the early pages that there are substantial differences between the versions. In most instances, these re-ordered statements were otherwise identical (excluding the global names changes).
3. Minor changes to the format of multi-line statements – For example, a three line statement is shortened to a two line statement by removing an extra line-return which changes only its appearance. The functional meaning of the statement is identical so the change is only cosmetic. There were 75 instances of this.
4. Addition and removal of non-functional lines in the code – For example, 200 blank lines were removed in the M version which changed the appearance but not the function. This is the primary cause of the difference in the total number of lines between the two versions and is obviously cosmetic.
5. Additionally, the M version added comments and section title lines which had no functional significance but did change the appearance. 109 lines are added titles, footnotes or comments which are non-functional.

It may be argued that additional commentary and improved cosmetics adds to the value of the source code because it enhances the understanding to the reader. This is true but it is not relevant to the assertions made – i.e., "… totally modified and rewrote …". It also has no bearing on the important issue of derivation. In other words, adding explanatory information to a published work does not change the ownership of the original.

Sincerely,

John Cosgrove, P.E.
Consulting Engineer

Graeme Shirley
Analyst

Exemplar -- Code Comparison

Description of the Analysis

The chief tool of the investigation was an Excel workbook which contains a set of spreadsheets which compare the two versions line by line. The M file contains 1178 lines, L 's 1469. The difference in the number of lines is primarily due to more than 200 extra blank lines that appear in L 's and two clearly defined sections that do not appear in M '.

Given that, it is straightforward to align every M line (save one) with a corresponding L line. Table 1 shows the correspondence. What is striking about the alignment is that in the first 100 lines (a series of Value statements), the order of the statements (which has no effect on the calculations) is shuffled considerably compared to the L order. However the order of the last 840 M lines is the same as the corresponding L lines.

The completion of aligning the two files allows a direct line-by-line comparison on the Excel spreadsheet. Of the 1178 M lines, 233 match exactly. Not an impressive total, until one notes that 20 variable names differ between the two files (a fact that was used in aligning the files). It takes less than 15 minutes to do a global change of "sex" to "gender," "rite_pas" to "passrt," and so forth. Now, of the 1178 lines, 929 are identical, and 109 are added titles, footnotes, or comments.

If one considers that "identical" means "identical as parsed by a computer", that means that character-by-character the lines must be identical. This includes leading, trailing, and contained spaces, the order of a series of comparisons, the point at which a long statement is broken over two or more lines, indentations, and other choices one makes when programming. Yet after only one edit (changing variable names) 929 of 1069 lines (excluding the added titles and footnotes) are identical, and the vast majority is in the same order.

The structure of the files is a series of statements, many of which are long, and broken into two or more lines. For example, the following statement (terminated with a semicolon) is three lines

```
ELSE IF (AGE EQ 11) AND (GENDER EQ 1) AND (TEST EQ 3) AND (SCORE GE 1 AND
SCORE LE 3)
THEN ST_END=2;
```

and (to the computer) differs from

```
ELSE IF (AGE EQ 11) AND (GENDER EQ 1) AND (TEST EQ 3) AND (SCORE GE 1 AND
SCORE LE 3)THEN ST_END=2;
```

although both are functionally the same.

Concatenating lines (75 instances) such as the above, which are split in L 's file, but not M ', and adjusting the amount of indentation on other lines, a short series of cosmetic changes can account for all but 23 lines of M version deriving directly from L 's.

Of the 23 remaining lines, 4 differ only in changing "fail" to "not pass;" 12 are due to adding three input fields (N_5, N_7, and N_9) to the input, thus changing the column numbers for other fields. (These variables only appear later in a statement which has been converted to a comment.) And one line differs because the location of the input file

is not the same. The remaining lines differ in the arguments on Data, Set, and Options lines.

To summarize, if one takes the L file, removes blank lines and comments (which includes old code converted to comments), shuffles 10 statements (4 or 5 lines each) in the first 100 lines, moves a few other blocks of lines, changes 20 variable names in an editor, removes 75 line returns to shorten statements, and changes some indentation, there is at least a 98% match with the M file. All of those listed changes, and most, if not all, the remainder, are purely cosmetic. One could not reasonably describe that as totally modified or rewritten.

Exemplar -- Code Comparison

Comparison Methods

I compared the M SAS file "PE Report SAS Commands.sas" to the L file "2Group2000.sas" in the following manner:

1) Copied the M file into the second column of an Excel spreadsheet ("PE Report")
2) Numbered each of the lines of the M file from 1 through 1178 in column 1 of its spreadsheet
3) Copied the L file into the second column of a second Excel spreadsheet in the same workbook as the M file (2Group2000)
4) Numbered each of the lines of the L file from 1 through 1469 in column 1 of its spreadsheet
5) Created a third spreadsheet ("Basic Comparison") which established a correspondence between each line of the M file and a line of the L file; the ordering is described generally in Table 1
6) Added a column to the Basic Comparison spreadsheet to check for matches of the 1178 lines to corresponding L lines; 233 match exactly
7) Copied the Basic Comparison spreadsheet to a fourth spreadsheet ("Renamed Variables") and changed the variable names in the L column according to Table 2; 929 lines match exactly and 109 are titles, footnotes or lines that were converted to comments in the M file
8) Copied the Renamed Variables spreadsheet to a fifth spreadsheet ("Breaks & Spaces") and concatenated statements (75 instances, plus two equations where a term was moved affecting 4 lines) that were split onto multiple lines on L , but on single lines on M , reordered 9 formats in first 100 lines, added spaces where only spaces differed between lines; 1046 lines match exactly, plus 109 titles, footnotes and comments

Exemplar -- Code Comparison

Line Number Alignment

M	L
1-3	1-3
4-7	24-27
8-11	52-55
12-15	19-22
16-24	56-64
25-29	47-51
30-33	42-45
34-37	33-36
38-43	5-10
44-51	65-72
52	74
53-55	11-13
56	73
57-58	75-76
59	
60	80
61-72	84-94
73-79	95-101
80-81	103-104
82-83	106-107
84-96	110-122
97-98	130-131
99-100	126-127
101-104	132-135
105	124
106-113	136-145
114-119	150-155
120-125	comment
126-214	197-298
215-302	400-500
303-337	157-195
338-1178	501-1469

Notes:

1) M lines 4-51 are initial format statements which are internally in a different order than L 's.
2) M line 59 does not correspond to any line in the L file.
3) The different position of M Line 105 relative to 101-104 may have an effect on the calculations.
4) A difference between the two files in the number of lines in a corresponding block is due to either embedded blank lines, concatenated lines, or added comments (Titles and Footnotes).

Exemplar -- Code Comparison

Global Names Changes

L	M
RITE_PAS	PASSRT
LEFT_PAS	PASSLF
SEX	GENDER
MILE_W	WALK
MILE_S	MILERUN
MILE_SS	
SSS_S	STRETCH
BACK_S	SIT_RCH
SCORE_S	ST_END
PACER_S	PACER
CURL_S	CURL_UP
TRUNK_S	TRK_LIFT
BMI_S	BODY
PER_FATS	FAT
DISTRICT	DISTRICT
C_NUM	CDS_C
D_NUM	CDS_D
SCHOOL	SCHOOL
S_NUM	CDS_S
LOCAL	LOCAL
	N_5
	N_7
	N_9
LNAME	NAME_L
FNAME	NAME_F
MNAME	NAME_M
BIRTHDTE	DOB