

Refactoring meets Spreadsheet Formulas

Sandro Badame, Danny Dig
 University of Illinois
 {badame1,dig}@illinois.edu

Abstract—The number of end-users who write spreadsheet programs is at least an order of magnitude larger than the number of trained programmers who write professional software. We studied a corpus of 3691 spreadsheets and we found that their formulas are riddled with the same smells that plague professional software: hardcoded constants, duplicated expressions, unnecessary complexity, and unsanitized input. These make spreadsheets difficult to read and expensive to maintain. Like automated refactoring in the object-oriented domain, spreadsheet refactoring can be transformative.

In this paper we present seven refactorings for spreadsheet formulas implemented in `REFBOOK`, a plugin for Microsoft Excel. To evaluate the usefulness of `REFBOOK`, we employed three kinds of empirical methods. First, we conducted a User Survey with 28 Excel users to find out whether they preferred the refactored formulas. Second, we conducted a Controlled Experiment with the same 28 participants to measure their productivity when doing manual refactorings. Third, we performed a Retrospective Case Study on the EUSES Spreadsheet Corpus with 3691 spreadsheets to determine how often we could apply the refactorings supported by `REFBOOK`. The results show: (i) users prefer the improved quality of refactored formulas, (ii) `REFBOOK` is faster and more reliable than manual refactoring, and (iii) the refactorings are widely applicable. On average `REFBOOK` is able to apply the refactorings in less than half the time that users performed the refactorings manually. 92.54% of users introduced errors or new smells into the spreadsheet or were unable to complete the task.

I. INTRODUCTION

The number of end-users who write spreadsheet programs is at least an order of magnitude larger than the number of trained programmers who write professional software [1]–[3]. Therefore the majority of programming is actually performed by users who do not consider themselves programmers. These spreadsheet users are referred to as end-users [4]. While these end-users are responsible for the maintenance and correctness of their spreadsheets, they have not been trained to develop software and often are not trained in the best practices for spreadsheet maintenance.

We have analyzed 3691 spreadsheets from the EUSES Spreadsheet Corpus [5] to determine the internal quality of the spreadsheets. We found that many formulas have *smells*, similar to those commonly found in professionally developed software. Some smells degrade the performance, others decrease readability, and others make it harder to change the table in the future. For example: 61% of formulas contain numerical constants that can be extracted. By consolidating all of the constant references in a sheet to a single place,

the formula’s readability and maintainability is increased. 13.66% of text columns are good candidates for conversion to a dropdown menu which reduces the possibility of typos occurring in a column and convey to a maintainer the acceptable values for a text column. 61% of formulas can be given descriptive names instead of using anonymous cell references, thus making it easier to understand formulas.

Although smelly formulas may correctly perform their tasks, they are difficult to maintain and can mask errors. Such errors have cost millions of dollars [4]. Researchers [6]–[18] have made continuous strides into finding and displaying errors and smells in spreadsheets. However, there is no work on the removal of smells from spreadsheet formulas.

In professional programming the removal of smells while preserving program behavior is called refactoring [19]. Refactoring is an important part of professional software development. Refactoring has revolutionized how programmers design software: it has enabled programmers to continuously explore the design space of large codebases, while preserving the existing behavior. Software has been found to have an inverse relationship between the number of applied refactorings and software defects [20]. Modern IDEs such as Eclipse, NetBeans, IntelliJ IDEA, or Visual Studio incorporate refactoring in their top-level menu and often compete on the basis of refactoring support.

Professional programmers have the support of refactoring tools. End-users, who are not even trained to maintain software, do not have any refactoring support. We propose to remove spreadsheet smells through the use of automated refactoring, analogous to the practice of object-oriented code.

There is a large number of refactorings for spreadsheet formulas that we could have implemented. However, we want to automate those that are frequently performed but cause frustration, and those that are infrequent but are difficult. We asked these questions to different end users.

We contacted the 750 members of the European Spreadsheet Risks Interest Group [21] that subscribe to the mailing list of professional spreadsheet users, and we exchanged several emails with their Chair, Patrick O’Beirne, the author of an influential paper [14] about best practices for spreadsheets. We also posted on two large forums that have been used by hundreds of thousands of users: the OpenOffice Calc forum [22] that has more than 200 active users online at any time, and the Excel forum [23] that has on average 4,000 active members online at any time. We also asked on the staff mailing list at the CS department at UIUC.

A

	A	B	C	D	E	F	G	H	I	J	K	L
1	Name	Apples	Oranges	Pinapples	Pears	Total Price	Sold Price	Fruits Sold	Remaining Fruits	Income	ROI	Favorite
2	Peter	3	1	18	4	13	14	14	12	1	0.0769231	Apples
3	John	6	15	5	8	17	20	20	14	3	0.1764706	Oranges
4	Sally	20	12	2	3	18.5	23	23	14	4.5	0.2432432	Pears
5	Marc	4	7	4	4	9.5	7	7	12	-2.5	-0.2631579	Apples
	STRING	NUMBER	NUMBER	NUMBER	NUMBER	(B5+C5+D5+E5)*0.5	NUMBER	G5/1	(B5+C5+D5+E5)-H5	G5-F5	J5/F5	STRING

↓

B

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Name	Apples	Oranges	Pinapples	Pears	Total Fruits	Total Price	Sold Price	Fruits Sold	Remaining Fruits	Income	ROI	Favorite	Constants	
2	Peter	3	1	18	4	26	13	14	14	12	1	0.076923	Apples	Purchase PPF	0.5
3	John	6	15	5	8	34	17	20	20	14	3	0.176471	Oranges	Sale PPF	1
4	Sally	20	12	2	3	37	18.5	23	23	14	4.5	0.243243	Pears		
5	Marc	4	7	4	4	19	9.5	7	7	12	-2.5	-0.26316	Apples		
	STRING	NUMBER	NUMBER	NUMBER	NUMBER	SUM(\$B5:\$E5)	(\$F5)*PurchasePPF	NUMBER	\$H5/SalePPF	\$F5-\$I5	\$H5-\$G5		STRING		
															IF(\$G5<=0,\$K5/\$G5,"Unknown")

Figure 1: Table before refactoring (A), and after refactoring (B). Below the table we show the formulas from row 5. Other formulas differ only in their row index.

Based on their input, we have implemented `REFBOOK`, a plugin for Microsoft Excel, that implements seven refactorings that reliably remove spreadsheet smells.

`REFBOOK` implements the following refactorings: `EXTRACT COLUMN`, `MAKE CELL CONSTANT`, `GUARD CALL`, `REPLACE AWKWARD FORMULA`, `STRING TO DROPDOWN`, `INTRODUCE CELL NAME`, and `EXTRACT LITERAL`. Each refactoring is specialized to remove a particular smell from a spreadsheet. These refactorings increase programmer productivity by performing the refactorings quickly and correctly. `EXTRACT COLUMN` breaks formulas into smaller components and can reduce the amount of code duplication that exists in spreadsheets. `MAKE CELL CONSTANT` makes formulas less error prone and more readable by rewriting the formula to contain `$`'s that signify that a particular cell or column is constant throughout a set of formulas. `GUARD CALL` rewrites a cell formula to have user defined behavior when an error condition occur. `REPLACE AWKWARD FORMULA` re-writes formulas using Excel's built-in functions (e.g., `SUM`) so that spreadsheets become more uniform and easier to understand. `STRING TO DROPDOWN` limits the number of allowed values for a cell to reduce the chance of a typo. `INTRODUCE CELL NAME` removes anonymous cells and replaces them with named cells. `EXTRACT LITERAL` removes "magic numbers" from formulas.

To evaluate the usefulness of `REFBOOK`, we employed three kinds of empirical evaluation methods. First, we conducted a User Survey with 28 Excel users to find out whether they preferred the refactored formulas. Second, we conducted a Controlled Experiment with the same 28 participants to measure their productivity when doing manual refactorings. Third, we performed a Retrospective Case Study on the EUSES Spreadsheet Corpus with 3691 spreadsheets.

This paper makes the following contributions:

- 1) To the best of our knowledge, we are the first to present refactorings in the domain of spreadsheet formulas.
- 2) We present the first refactoring tool for spreadsheet formulas, `REFBOOK`, implemented as a plugin for Excel. `REFBOOK` currently supports seven refactorings. A demo can be seen at: <http://www.youtube.com/watch?v=wGIu6Muvd8I>

3) Our three-way evaluation reveals:

- (i) For four out of the seven refactorings users preferred the refactored output. Thus the refactorings *improve spreadsheet quality*.
- (ii) On average `REFBOOK` is able to apply the refactorings in less than half the time that users performed the refactorings manually. Thus `REFBOOK` *improves programmer productivity*.
- (iii) 92.54% of users asked to perform the same refactorings introduced errors into the spreadsheet or where unable to complete the task. `REFBOOK` makes it easier to apply the refactorings correctly, thus it is *more reliable*.
- (iv) The refactorings can be applied to many of the formulas contained in spreadsheets. Thus `REFBOOK` is *applicable*.

II. MOTIVATING EXAMPLE

To illustrate the kinds of refactorings applicable to spreadsheets, we will use the table shown in Figure 1(A). This table tabulates data from a warehouse where four salespersons purchased fruits and resold them for a profit. Each row tabulates the data for one salesperson. Under the table we show the kind of each column and the formulas they compute. Notice that we only show the formulas as they would appear in row 5 (thus referring to cells from row 5), but the formulas for the other rows refer to their respective cells.

The table contains twelve columns: six columns contain literals, and six columns contain formulas that compute on the other columns. Now we briefly explain each column.

Column A is a text column with the names of the sellers.

Columns B-E are numerical columns that hold the number of each type of fruit that each salesperson had purchased.

Column F is a formula column that computes the price that each seller paid for their fruits. It performs two calculations: sum the number of fruits purchased by each person, then multiply it by \$0.50, the purchase price per fruit.

Column G is a numerical column that holds the amount of money that each seller collected from selling their fruits.

Column H is a formula column that computes the number of fruits sold: it divides the `Sold Price` by \$1, the resale price per fruit.

Column I is a formula column that computes the remaining

fruits that each salesperson still has. This column performs two calculations: sum the total number of fruits purchased by that person, then subtract their number of fruits sold.

Column J is a formula column that computes `Income` as the difference between `Sold Price` and `Total Price`.

Column K is a formula column that computes the return on investment from `Income` and `Total Price`.

Column L is a text column containing the favorite fruit as reported by the salesperson. This column should only contain the names of real fruits, not arbitrary text.

There are several “smells” in this table. Some smells degrade the performance, others decrease readability, and others make it harder to change the table in the future.

Take for example the expression `B5+C5+D5+E5` which computes the total number of sold fruits. First, this can become more readable if it was replaced with the built-in `SUM` function (i.e., `SUM(B5:E5)`). Second, this expression is calculated twice, once in the `Total Price` column and then again in the `Remaining Fruits` column. Given that Excel does not cache expression results, but instead does cache cell results, this wastes CPU cycles. Third, since this expression is duplicated between the `Total Price` column and the `Remaining Fruits` column, a future change request like introducing a new kind of fruit and its afferent column, requires changing the expression in the two columns. Like in the case of professional programming, duplication increases the maintenance effort.

Also notice that the table contains two constants, `0.5` and `1`. First, constants make the formulas unreadable: another co-worker who inspects the table will have to guess what is the meaning of these constants (i.e., purchase price and resale price). Second, constants make it tedious to perform maintenance tasks: if we wanted to change the purchase and resale price, we will have to manually find and update 8 cells. Performing a find-and-replace for `1` will erroneously update the cell `C2`, which just happens to have value `1`.

Also notice that the `Favorite` column can contain any arbitrary text, even the ones that are not fruit names, for example by a typo (e.g., “Apples”). This affects the readability of the column for humans. This is even worse for other automated tools: macros or other programs that read such erroneous values won’t work.

Also notice that all formulas that refer to static cells use the “fixed column” format. For example, the formula in `H5` refers to the static cell `G5`. Adding the `$` can make cell references more resistant to errors when the spreadsheet is modified, e.g., when the formula is dragged down a column.

Dragging a formula in Excel copies the formula into the adjacent cell and changes the cell references in the formula to reference the new row or column that the cell was dragged into. For example, dragging `A1` down one row will insert the formula `A2` into the new row. While the ability to drag a formula is very useful in practice, it leads to smelly and possibly erroneous spreadsheets when not used carefully.

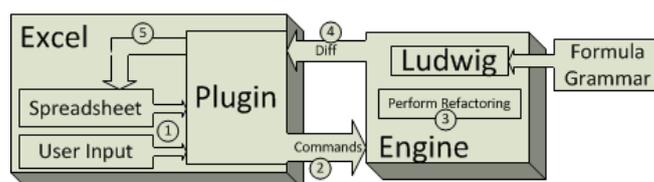


Figure 2: Architecture of REFBOOK.

Bugs can occur when dragging a formula that contains a reference to a cell that is constant throughout the entire table. Dragging this formula will also increment the reference to the constant cell. This means that the dragged formulas will refer to a different cell than the one intended by the user.

The `$` signs also serve as a form of documentation about the formula: the `$` signs highlight which cell references are constant throughout the entire column of formulas.

Figure 1(B) shows the same table after we applied several refactorings. We added the `Total Fruits` column to hold and intermediate calculation of total number of purchased fruits. The `Total Price` and `Remaining Fruits` formulas are modified to reference the `Total Fruits` column.

We moved the `0.5` and `1` constants from `Total Price` and `Fruits Sold` columns into their own cell and named them `PurchasePPF` and `SoldPPF` respectively.

We changed the `ROI` column to sanitize its input against a division by `0`. If the number of sold fruits was `0` then the `ROI` would display a “divide by zero” error. With the guarded formula, the “Unknown” text is displayed. The advantage is that it notifies the reader that the formula was written with the error case in mind, and that the error case is handled.

We changed the `Favorite` column to use a dropdown menu for valid values instead of arbitrary text. Using a dropdown also makes the user aware that there are a fixed number of values that are allowed in the column.

Across all of the formula cells, we updated the cell references to use the `$` on their column identifier.

REFBOOK supports all the refactorings performed on the table from Fig. 1(B).

III. HIGH LEVEL OVERVIEW OF REFBOOK

A. Typical use of REFBOOK

Users interact with REFBOOK from within Excel. We describe a typical use of REFBOOK using our motivating example from Figure 1: replace the formulas in the `Total Price` column with formulas that contain constant cell references. To perform this refactoring the user selects cells `F2-F5` from the `Total Price` column, right-clicks on the selection to bring up the context menu (which shows REFBOOK on the top), then selects the `MAKE CELL CONSTANT` option in the menu. REFBOOK then replaces the highlighted formulas with formulas that contain the `$` sign.

B. Life-cycle of a Refactoring

REFBOOK’s architecture consists of three major components: Excel plugin, Ludwig [24], and the refactoring engine. The

Excel plugin is the front-end. Users only interact with the Excel plugin component of REFBOOK. The Excel plugin creates a separate process for the back-end, Ludwig and the refactoring engine. The back-end calculates the corresponding changes for each refactoring, and sends them to the Excel plugin that applies them on the spreadsheet.

Ludwig [24] is an off-the-shelf component that given a grammar for a language (e.g., the grammar for Excel formulas – see [25]) generates a Java library for parsing that grammar. The Abstract Syntax Tree (AST) generated by Ludwig supports manipulation of the AST while preserving the formatting for the remainder of the formula. The refactoring engine is a Java application that takes as input the name of the refactoring to perform, the table that is the target of the refactoring, and the user input. It performs the AST transformations, e.g., adding, removing, updating cells. Then it outputs commands for the Excel plugin to perform on the Excel spreadsheet. The output of the refactoring engine is an ordered list comprised of commands of this kind:

- INSERTCOLUMN columnIndex
- INSERTROW rowIndex
- SETCELL columnIndex, rowIndex, content
- NAMECELL columnIndex, rowIndex, name

The major advantage of this 3-tier architecture is that REFBOOK is extensible to other spreadsheet tools, beyond Microsoft Excel. If we were to implement refactoring support for OpenOffice Calc [26] then we would have to reimplement only the Excel plugin portion of REFBOOK.

IV. REFACTORINGS

A. Anatomy of a Spreadsheet

The terminology we will use is derived from Excel’s terminology. A *Workbook* is a single file that contains multiple *Sheets*. A sheet can only belong to one *Workbook*. A *Sheet* is a named two dimensional array of *Cells*.

Cells are indexed in a *Sheet* either by row and column, or by a user-defined name. A *Cell*’s column is represented as series of letters that range between A and ZZ. A *Cell*’s row is represented as an integer greater than 0. *Cells* can be named or anonymous. The condition for a valid user-defined name is that the given name cannot be interpreted as a valid cell index (e.g., A1 is not a valid cell name) or contain spaces. We will refer to a cell that has not been given a user-defined name as an anonymous cell.

Cells can contain a value that is one of three types: Number, Text, or Formula. REFBOOK parses a formula into AST nodes using the grammar from our technical report [25].

B. Definitions

It is a common practice for a user to define one formula in one cell and then drag it down a column or across a row. However, the user only created one single formula, the dragged cells only differ by the cell references. Many of our refactorings check whether the user selected cells that

belong to consistent formulas (defined below), e.g., dragged formulas.

Definition 1. *Consistent formulas are formulas that have the same AST shapes and their corresponding formula AST nodes have the same names.*

Definition 2. *Two formulas have the same shape if their ASTs are isomorphic, i.e., the ASTs contain the same number of AST nodes, the corresponding nodes have the same type, and the nodes form the same structure.*

Definition 3. *Distinct formulas are formulas that are not consistent.*

Procedure 1 presents the pseudocode for REFBOOK’s implementation of ISCONSISTENT. This procedure takes as input two AST nodes, and it returns a boolean value. Notice that if the nodes contain other child AST nodes, the procedure further checks the children for consistency.

Procedure 1 ISCONSISTENT

```

function ISCONSISTENT(node1, node2)
  if node1.type! = node2.type then
    return False
  end if
  if node1.type == Function then
    if node1.functionName! = node2.functionName then
      return False
    end if
  end if
  if node1.children.count! = node2.children.count then
    return False
  end if
  for i = 0..(node1.children.count - 1) do
    if !isConsistent(node1.children[i], node2.children[i])
  then
    return False
  end if
  end for
  return True
end function

```

Now we describe each of the seven refactorings supported by REFBOOK. For each refactoring we present an example of the transformation, then we describe in plain text what the refactoring does, then we provide pseudo-code for the algorithm. Here we show the input, the preconditions, and the transformation. We use the same style of behavior-preservation introduced by Opdyke [27], namely we guarantee that the refactored spreadsheet computes the same values as the original spreadsheet when the preconditions are true.

We define one more notation used in the pseudocode of the refactorings: *formula.cellReferences* denotes the collection of *CellReference* nodes that the formula’s AST contains.

C. EXTRACT COLUMN

Example: In the motivating example (Fig. 1), we apply the refactoring to the *Total Price* column. It extracts the expression: (B5+C5+D5+E5) from *Total Price* and *Remaining Fruits* into a new cell, F5. The new column will contain formulas like B5+C5+D5+E5. The *Total Price* column will hold F5*.5 and the *Remaining Fruits* will hold F5-H5.

Description: The user selects a column to extract from and a subexpression from the column to be extracted into a new column. REFBOOK moves the selected column one position to the right. Then it updates the cell references in the table to refer to the new cell positions after the movement. It places the extracted subexpression into every cell of the newly created column. Then, it finds all instances of the subexpression in the table and replaces them with a reference to the corresponding cell in the new column.

Now we briefly describe the pseudocode in Procedure 2. After checking the precondition to ensure that formulas in the selected column are consistent, REFBOOK inserts a new column to the left of the user-selected column. Now we explain this in the `insertColumn` function. First, it shifts one column to the right all cells in the selected column and its succeeding columns. Second, it updates all cells references in the sheet that refer to any of the cells that we shifted.

Returning back to the main procedure, REFBOOK iterates through the cells of the newly created column. It populates this column with the updated sub-expressions. Function `computeRowExpr` computes the appropriate subexpression for each row (e.g., $(B6+C6+D6+E6)$ for row 6). Next, REFBOOK searches the entire row and replaces all references to that subexpression with references to the new cell. Once all of the replacements for that row are done, REFBOOK populates the new cell's formula with the updated subexpression.

Procedure 2 EXTRACT COLUMN

Input: sheet, userCell, expr
Preconditions: $\forall f1, f2 \in \text{sheet}[\text{userCell.column}], \text{isConsistent}(f1, f2)$

```

function EXTRACTCOLUMN
    newColumn = insertColumn(sheet, userCell.column)
    for all cell ∈ newColumn do
        rowExpr = computeRowExpr(expr, userCell.row, cell.row)
        for all c ∈ cell.row | c.formula.contains(rowExpr) do
            c.formula.stringReplace(rowExpr,
                newColumn + cell.row)
        end for
        cell.formula = " = " + rowExpr
    end for
end function

function INSERTCOLUMN(sheet, c)
    for all cell ∈ sheet[cell.column] >= c do
        sheet[cell.column + 1][cell.row].formula = cell.formula
    end for
    for all cell ∈ sheet do
        for all cellRef ∈ cell.formula.cellReferences do
            if cellRef.sheet = sheet && cellRef.column >= c then
                cellRef.column += 1
            end if
        end for
    end for
end function

function COMPUTEROWEXPR(oldExpr, oldRow, newRow)
    expr = copy(oldExpr)
    for all cell ∈ expr.cellReferences do
        if cell.row = oldRow then
            cell.row = newRow
        end if
    end for
    return expr
end function

```

D. MAKE CELL CONSTANT

Example: In the motivating example in Fig 1, we apply the refactoring to all of the formula columns: Total Price, Fruits Sold, and Remaining Fruits. REFBOOK converts the column formulas from $(B5+C5+D5+E5)*0.5$, $G5/1$ and $(B5+C5+D5+E5)-H5$ to $(\$B5+\$C5+\$D5+\$E5)*0.5$, $\$G5/1$, and $(\$B5+\$C5+\$D5+\$E5)-\$H5$ respectively.

Description: The user selects a whole column. REFBOOK first determines whether any of the cell references can be made constant. It uses the shape of the first formula as the model for all the other formulas in the selection. Then it compares corresponding cell references between pairs of selected formulas and it determines which cell references do not change (the references with \$ prefixes). Due to space limitations, we present the pseudocode in [25].

E. GUARD CALL

Example: In the motivating example we apply the refactoring to the ROI column. The user supplied the error expression "Unknown". GUARD CALL converted $J5/F5$ to $\text{IF}(G5 < > 9, K5/G5, \text{"Unknown"})$.

Description: The user selects a formula cell and also provides an expression to be supplied as the error message. REFBOOK searches for a division operator, and replaces it with a conditional IF, where the condition checks whether the denominator is different than zero, the then branch performs the division, and the else branch displays the error message.

Procedure 3 GUARD CALL

Input: formula, errMsg
Preconditions: $\exists "/" \in \text{formula}$
 errMsg.isValidFormula

```

function GUARDCALL
    binaryOps = formula.collectAll(AnotherExpression)
    for all n ∈ binaryOps | n.Operator = "/" do
        guard = "IF(" + n.Expression + " <> 0, " + n.Parent +
            ", " + errMsg + ")"
        n.parent.ExpressionPrimitive = guard
    end for
end function

```

F. REPLACE AWKWARD FORMULA

Example: In the motivating example, we apply the refactoring to the Total Price and Remaining Fruits fruits column converting them from $(B5+C5+D5+E5)*0.5$ and $(B5+C5+D5+E5)-H5$ to $\text{SUM}(B5:E5)*0.5$ and $\text{SUM}(B5:E5)-H5$ respectively.

Description: The user selects a formula and REFBOOK first searches for expressions containing the + or * operator, and at least four operands of consecutive cells. REFBOOK replaces such long chains with a single `SUM(<<range>>)` or `PRODUCT(<<range>>)` function.

G. STRING TO DROPDOWN

Example: In our motivating example, we apply this refactoring to the Favorite column. The set of valid entries consists of: Apples, Oranges, and Pears.

Procedure 4 REPLACE AWKWARD FORMULA

Input: formula
Preconditions: $\exists \text{“+” or “*”} \in \text{formula}$
 $\exists \{\text{cellRef}\} \in \text{formula} \mid \{\text{cellRef}\}.cardinality > 3$
function REPLACEAWKWARD
 $original = \text{formula}$
 $refactored = \text{attempt}(\text{formula})$
 while $original \neq refactored$ **do**
 $original = refactored$
 $refactored = \text{attempt}(original)$
 end while
end function
function ATTEMPT(expr)
 $awkwardAST = \text{getAwkwardASTNode}(expr)$
 if $awkwardAST \neq \text{NONE}$ **then**
 $expr.\text{replace}(awkwardAST, \text{fixedAST})$
 end if **return** expr
end function

Description: The user selects a textual column. REFBOOK finds all the unique text entries in the column. REFBOOK attaches dropdown menus to each cell in the column.

Procedure 5 STRING TO DROPDOWN

Input: column
Preconditions: $\forall \text{cell} \in \text{column}, \text{cell}.isTextual$
function STRING TO DROPDOWN
 $choices = \{x \in \text{column} \mid x \notin \text{choices}\}$
 for all $\text{cell} \in \text{column}$ **do**
 $\text{cell}.DropdownOptions = \text{choices}$
 end for
end function

STRING TO DROPDOWN assumes that the user-selected column does not contain any errors. Specifically if typos exist, they also populate the values in the dropdown menu.

H. INTRODUCE CELL NAME

Example: In our second table from the motivating example, we applied this refactoring to the cell O2, and we named it PurchasePPF. The formula in G5, F5*PurchasePPF, uses the new name.

Excel’s “Search and Replace” is not a sound method to use to replace all anonymous cell references in formulas with the named reference. For example, if the text A1 was in a text literal then it would be replaced. If A1 was referenced as \$A\$1 “Search and Replace” misses this reference. Without REFBOOK the end-user programmer would be forced to inspect each cell in the table for correctness. REFBOOK reliably and correctly finds all references to the cell.

Description: The user selects a cell, and provides a name. REFBOOK checks whether the new name is not in use, and defines the name. REFBOOK searches in the entire table for references to the anonymous selected cell, and updates them to the named cell.

I. EXTRACT LITERAL

Example: In the motivating example, we apply the refactoring to the Total Price and Fruits Sold columns. Each of these columns have a “magic number” (0.5 and 1).

Procedure 6 INTRODUCE CELL NAME

Input: anonCell, name, sheet
Preconditions: $\text{name}.isValid$
function INTRODUCE NAME
 $anonLoc = \text{location}(anonCell)$
 $\text{sheet}.defineName(anonCell, \text{name})$
 for all $\text{cell} \in \text{sheet}$ **do**
 for all $\text{cellRef} \in \text{cell}.Formula$ **do**
 if $\text{location}(\text{cellRef}) = anonLoc$ **then**
 $\text{cell}.Formula.\text{replace}(\text{cellRef}, \text{name})$
 end if
 end for
 end for
end function
function LOCATION(cell)
 return $(\text{cell}.Workbook, \text{cell}.Sheet, \text{cell}.Column, \text{cell}.Row)$
end function

Description: User selects a formula. She also selects the actual literal value from that formula to be extracted and provides a name for the cell. REFBOOK checks whether the new name is not in use. Then REFBOOK finds an empty cell that it moves the selected literal into, names the new cell. REFBOOK determines which AST node was selected to be extracted. Then REFBOOK searches all cells in the table for formulas that are consistent with the user-selected cell, then in all such consistent cell formulas, the AST node that was selected is replaced with a reference to the named cell.

Procedure 7 uses the function `getPathTo` to find the path from the root node of the formula’s AST to the `Primitive` that contains the literal node. `getPathTo` helps avoid the case of a literal in a different context being replaced.

Procedure 7 EXTRACT LITERAL

Input: literal, cellName, cell
Preconditions: $\text{cellName}.isValid$
function EXTRACTLITERAL
 $\text{sheet} = \text{cell}.sheet$
 $\text{path} = \text{getPathTo}(\text{cell}.formula, \text{literal})$
 $\text{sheet}.defineName(\text{sheet}.getUnusedCell(), \text{cellName})$
 $\text{sheet}[\text{cellName}] = \text{literal}$
 for all $c \in \text{sheet} \mid \text{isConsistent}(c.formula, c.formula)$ **do**
 $\text{node} = c.formula.\text{nodeAtPath}(\text{path})$
 $c.formula.\text{replace}(\text{node}, \text{cellReferenceTo}(\text{cellName}))$
 end for
end function

V. EVALUATION

We evaluate the usefulness of the proposed refactorings by answering four research questions:

- **Q1:** Do the refactorings improve the spreadsheets quality?
- **Q2:** Can REFBOOK make the refactoring process more reliable?
- **Q3:** Do the refactorings improve programmer productivity?
- **Q4:** Are the refactorings applicable?

All these questions address the higher level question “Is REFBOOK useful?” from different angles. Quality measures whether the users find the refactored formulas more readable. Reliability ensures that the runtime behavior is not modified and the transformation does not introduce more smells.

Productivity measures whether automation saves human time. Applicability measures how many formulas in real-world spreadsheets can be directly transformed.

A. Methodology

To answer these questions, we employed three different empirical techniques. (i) To assess whether users preferred the refactored formulas, we conducted an online User Survey with 28 Excel users. (ii) To measure refactoring reliability and user productivity when performing manual refactorings, we conducted a Controlled Experiment with the same 28 users. In order to ensure discretion, each participant responded to the survey and performed the change tasks in their own environment. (iii) To determine how often we could apply the refactorings supported by REFBOOK, we performed a Retrospective Case Study on the EUSES Spreadsheet Corpus [5] with 3691 spreadsheets.

Recruitment: To recruit participants, we advertised to students in the University of Illinois CS105 course. This course is attended by students enrolled in the Business department. In this course students learn how to use Excel for business-related purpose. The participation in the survey was voluntary, and it did not have any relationship with the course evaluation. The successful completion of the survey was rewarded with a \$5 Amazon giftcard.

Out of the 500 enrolled students, 28 responded to our call. We asked three questions about their experience with Excel. Figure 3 shows the demographics of our participants. Notice that two-thirds of the participants claimed to have more than two years experience with Excel. All our participants responded within 24 hours from our post.

1) **User Survey:** Each participant in the User Survey used an Excel document, consisting of 7 sections. Each section focuses on a single particular smell. Each section has two tables: one table that contains the smell and one table where we have previously used REFBOOK to remove the smell through one of our refactorings.

For each section, we asked two questions about the tables. The first question was a “filter”: we asked a technical question that revealed whether the participant studied the two tables and understood the differences between them. The second question asked which table they would prefer to work with. To eliminate the confounding effect, we randomized the order of appearance between smelly and non-smelly tables.

2) **Controlled Experiment:** Each participant used an Excel document. The tables contained data about the orchard warehouse, similar with the example shown in our motivating example from Fig. 1. The document contained 7 tasks. Each task has a smelly table, a set of instructions on what to change in the table, a “Start” button, and a “Task Complete” button. Before the “Start” button is pressed, the spreadsheet is read-only. During this time the participant is free to inspect and become familiar with the table, the task that she will perform, and a short, optional tutorial that we designed to

present the Excel features she might use.

Once the participant has familiarized with the task and the table, she can press the “Start” button. When a participant presses the “Start” button, the table becomes editable, and the participant can perform the changes. A timer records the time taken to perform the changes. When a participant completes the task, she presses the “Task Complete” button which stops the timer and moves her onto the next task.

After we received their online submissions, we processed each document to record the time it took participants, and whether they performed the tasks correctly. We also applied ourselves REFBOOK to complete the same tasks, and then we compared our results with the participants’ results.

3) **Retrospective Case Study:** To determine the applicability of our refactorings we analyzed the EUSES Spreadsheet Corpus’s 3691 spreadsheets to find out how many formulas have smells that can be fixed by our refactorings. We chose the EUSES Spreadsheet Corpus because it is regarded as the most mature, representative corpus of spreadsheets. At least 13 published papers [5] have used it to draw conclusions about spreadsheet programming. This corpus contains 206355 tables and 495578 distinct formulas.

First, we had written a tool to find the tables in all the spreadsheets. In real-world spreadsheets, tables are (i) often surrounded by documentation, (ii) do not begin at the top and left-most cell, and (iii) multiple tables are scattered throughout the spreadsheet. Our tool parses the corpus spreadsheets using the Apache POI [28] Java library. Due to limitations in the Apache POI [28] library, our tool could not parse 234 spreadsheets of the 3925, so we retained 3691. To find the individual tables that exist within a sheet, we used the algorithm described in [16].

Then we implemented a collector tool, customized for our refactorings. The collector calculates how many cells in each table manifest a particular smell that can be fixed by a particular kind of refactoring.

B. Results

Quality: Table I shows for each refactoring kind, how many participants preferred the smelly or the refactored formulas.

For 4 out of the 7 refactorings, the majority of participants preferred the refactored formulas. For one refactoring, STRING TO DROPDOWN, the majority did not have any preference. For 2 refactorings, the majority preferred the smelly formulas. For the INTRODUCE CELL NAME, the participants preferred the table that contained the anonymous cells. Since 82.61% of the participants were able to correctly answer the filter question about the table, we are confident that they understood the table. This corroborates another study [29] where end-users working with Yahoo pipes preferred seeing all the pipes at once instead of abstracting functionality to another pipe.

When judging whether the refactorings improve the readability for end users, we assume that their opinion reflects

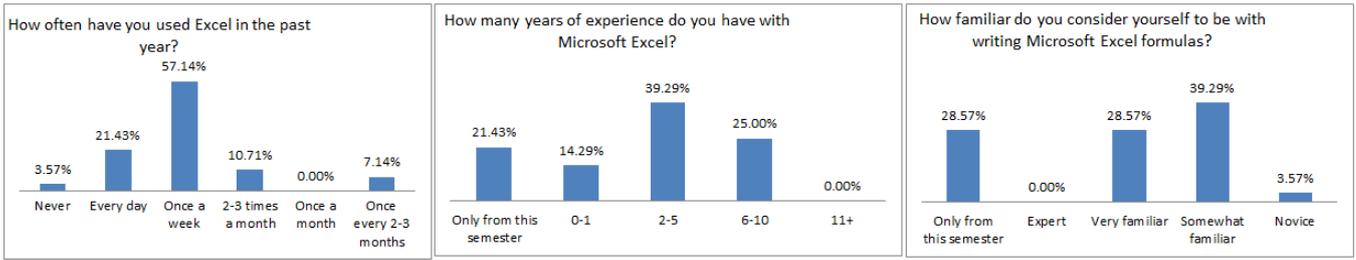


Figure 3: Demographics of our 28 participants.

Refactoring Kind	Prefer Smelly	Prefer Refactor	No Pref.	No Resp.	Pass Filter
ExtractColumn	17.39%	47.83%	21.74%	13.04%	73.91%
MakeCellConstant	21.74%	52.17%	13.04%	13.04%	60.87%
StringToDropdown	8.70%	21.74%	56.52%	13.04%	34.78%
ReplaceAwkward	4.35%	52.17%	26.09%	17.39%	78.26%
GuardCall	47.83%	13.04%	26.09%	13.04%	78.26%
IntroduceName	52.17%	4.35%	26.09%	17.39%	82.61%
ExtractLiteral	17.39%	60.87%	8.70%	13.04%	69.57%

Table I: Preferences of users toward formulas.

the quality of the formulas. Others [29] have used the same technique when judging the quality of end-user code.

Reliability: The second column in Table II shows how many of our participants submitted a solution that contained at least one fault. A fault is a semantical error (i.e., the changed formula computes the wrong value) or a smell (i.e., the original smell was not corrected). Next to each percentage of faults there are two numbers. The first number is the percentage of users that committed a semantical error. The second number is the percentage of users that missed an opportunity to perform a transformation that REFBOOK would have performed.

The overall majority of the 28 participants submitted solutions with at least one fault. For one refactoring, EXTRACT COLUMN, all submitted solutions had faults. An example of semantical error is when some participants extracted the wrong subexpression into a new column. For the example in Fig. 1, they had to extract the $B5+C5+D5+E5$ from column F containing the formula $(B5+C5+D5+E5)*0.5$, into a new column G. However, they copied the wrong expression $(B5+C5+D5)$ into a new column G, and replaced the original formula with $G5*0.5$. Unfortunately, this new formula computes a different value (due to missing E5).

An example of missed opportunity, many participants copied the subformula into a new column, but did not remove the duplication between the newly introduced column and the old column. That is, the new column is never referenced from the old column. For the example in Fig. 1, from column F containing the formula $(B5+C5+D5+E5)*0.5$ they copied the expression $(B5+C5+D5+E5)$ into a new column G, but did not replace the original formula with $G5*0.5$.

In contrast, REFBOOK performs all refactorings correctly.

Productivity: The third and fourth columns in Table II show the average time that it took our 28 participants to perform the manual refactorings. The fifth column shows the time

Refactoring Kind	% Faulty Submissions	Manual Time[sec]	Std. Dev.	REFBOOK Time[sec]
ExtractColumn	100 (26+74)	36	18	16
MakeCellConstant	30 (26+4)	37	17	09
StringToDropdown	82 (82+0)	217	169	09
ReplaceAwkward	47 (47+0)	68	42	22
GuardCall	82 (82+0)	67	28	31
IntroduceName	82 (56+26)	79	50	30
ExtractLiteral	91 (34+57)	42	23	18

Table II: Safety & Productivity of manual vs. automated refactorings.

we took to perform the same refactorings with REFBOOK.

Notice that the time that our Excel macro records for the participants includes both the selection of cells and the actual change. In the REFBOOK's time we also report the time to select cells and to apply REFBOOK (though REFBOOK applies the refactoring in less than 3 seconds).

The table shows that performing the refactoring with REFBOOK is faster than performing it manually, the improvements ranging from 2.2x to 24x. This is a conservative lower bound; we expect the productivity difference to be even more dramatic in practice. First, real-world tables contain more rows than the 15 rows in our controlled experiment. Second, the faults committed by the participants were typically errors of omission: many participants had applied the refactoring incompletely. Had they applied the complete refactoring, this would have taken them even more time.

Applicability: Based on the EUSES Spreadsheet Corpus, we present the applicability of individual refactorings.

There are many formulas in a spreadsheet that are dragged down a column or across a row. However, the user only created one single formula, the rest only differ by the cell references. If we took these dragged cells into account, our results will be skewed by the amount of dragged cells that exist in each table. To prevent this, we define and compute metrics only over *distinct formulas* (defined in Section IV).

EXTRACT COLUMN. First, we measured the formula complexity. We introduce the following metric to measure a formula's complexity: the sum of the number of binary operators and function calls that a formula contains. For example:

```
IF(G5 <> 0, K5/G5, "Unknown")
```

has a complexity of 3 (one function call, i.e., IF, plus two binary operators, i.e., not equals and division). A formula that contains only a single reference to another cell, number, or text, has a complexity of 0.

We found that 10.1% of distinct formulas have a complexity of 0, 57.71% have a complexity of 1, 18.49% have a complexity of 2, 11.97% have a complexity of 3, 1.73% have a complexity greater than 3. Formulas that have complexity larger than 1, (32.19% of all distinct formulas), are candidates for the `EXTRACT COLUMN`, which reduces the complexity of a formula by breaking it into smaller sub-formulas.

We also compute the amount of duplication that exists between distinct formulas in a table. We calculate the amount of duplication by counting the number of times an AST node is repeated in a table. 72.89% of the formulas contain no duplication. The remaining 27.11% contain some amount of duplication, thus are candidates for `EXTRACT COLUMN`.

MAKE CELL CONSTANT We apply this refactoring on every distinct formula and recorded the number of cell references that were successfully made constant. We found that 23.28% of the formulas did not change when we applied the `MAKE CELL CONSTANT` refactoring, 9.03% of the formulas had a single `$` prefix added to a cell reference, 19.35% of the formulas had two places where `$` was added to cell references (e.g., `A5`), 3.24% of the formulas had three `$` added to cell references, 32.64% of the formulas had four `$` added to cell references (e.g., `A5 + B5`).

GUARD CALL We found that `IFERROR`, `ISBLANK`, and `ISNUMBER` are among the top 10% most used Excel functions. This shows that explicit error handling is a popular technique.

REPLACE AWKWARD FORMULA We found that 15.73% of all distinct formulas use the `SUM` function. This shows that end-users understand it and like to use it.

STRING TO DROPDOWN We computed the number of duplicated entries that exist in a column of text values. We found that 86.34% of the text columns had no duplication of text values, 6.99% of the text columns repeated up to 50% of the text entries, and 6.67% repeated between 51% and 99% of their text entries. 95% of the text columns with 51% or more cells repeated can be converted into a dropdown menu with 10 or less different choices. We found that 13.66% of text columns are strong candidates for `STRING TO DROPDOWN`.

INTRODUCE CELL NAME We found that 61% of formula columns refer to at least one common anonymous cell that is a good candidate for being named. For example, column `C` contains formulas of the type: `A1+B0`, `A2+B0`, `A3+B0`; cell `B0` is a good candidate for `INTRODUCE CELL NAME`. We also found that in such columns, on average 2.21 cells could be named.

EXTRACT LITERAL Among formula columns, we found that 61% of them referred to the same numerical constants. For example, column `D` contains formulas of the type: `C1+12`, `C2+12`, `C3+12`; constant `12` can be extracted into a separate cell. Of the columns where `EXTRACT LITERAL` can be applied, on average 2.09 numerical constants can be extracted. We also found that 0.06% of formula columns referred to the same string literal. From these columns, on average 1.84 string literals can be extracted.

VI. DISCUSSION

Future Extensions: There are new refactorings that we could implement, some being trivial extensions of the current refactorings. For example, for implementing `EXTRACT ROW`, we could use the same pseudocode we used for `EXTRACT COLUMN` in Procedure 2 where we swap the columns with rows and vice-versa. Additionally, we could also implement refactorings that span multiple sheets

Also, we could make our prototype implementation more robust. For example, it currently does not check whether the user selects a subexpression that transcends the boundary of operator precedence. A user could select `2+3` from the `6*2+3` formula and apply `EXTRACT COLUMN`. An industrial-strength implementation should raise a warning that the new formula will compute value 30 instead of 15.

Threats to Validity: One could raise objections to our empirical findings based on the distribution of participants. 32% of our participants are self-reported novice Excel users. If we had more novices, they might have preferred the smelly version of the formulas, thus influence the outcome of our findings. However, we believe proficient Excel users can provide more valuable feedback about spreadsheets quality, similarly with how proficient Java programmers provide more valuable feedback for the Eclipse refactoring engine.

A very useful validation of refactorings could measure whether users find the refactored formulas easier to maintain. However, this is very hard in a study with students, because they never maintain their formulas.

VII. RELATED WORK

Erwig [12] proposes to apply techniques and tools from professional programming to end-users. Erwig specifically advocates for better error reporting, debuggers and static type checking [30]. He does not mention applying the refactoring techniques from professional programming.

Guidelines for creating clean, consistent and non-smelly spreadsheets have been proposed by others [13], [14].

`PUP` [17] and `ASAP Utilities` [18] are tools that add many features to Excel including some basic formula manipulation. The “Error Condition Wizard” in `PUP` and “Custom Error Message” in `ASAP Utilities` are similar in spirit to our `GUARD CALL` refactoring. The major difference is that these tools do not let users type in arbitrary expressions to be executed in the else branch, but they limit the type to Strings, unlike `REFBOOK` that allows any arbitrary expression. Also, our `GUARD CALL` infers the check for erroneous behavior, it does not react to an already existing error. This gives users more flexibility to define the action that should be taken for bad input.

`ASAP Utilities` includes a “Change formula reference style” operation and Excel has a feature that adds `$` sign to cell references that is similar in spirit to our `MAKE CELL CONSTANT`. However, they cycle through the selected cells and blindly add `$` to every single cell reference. Our `MAKE CELL CONSTANT` is different from the above alternatives because it intelligently

determines which cells should be made constant based on their usage in similar formulas from the user selection.

“What You See is What You Test” (WYSIWYT) [7] is a testing tool that helps end users find bugs in their spreadsheets. WYSIWYT estimates a cell’s correctness based on user input. WYSIWYT does not offer any automation to correct cells that are found to be incorrect.

Cunha et al. [8] detect data in spreadsheets that are outliers from the typical entries. These outliers are referred to as being “smelly”. Their work does not apply to finding spreadsheet formula smells. REFBOOK removes smells from formula cells not from data cells.

Hermans et al. [9]–[11] implemented tools to detect spreadsheet smells and visualize them. Their work is complementary to ours, since it focuses on making spreadsheet smells apparent through visualizations but does not support removing the smells. Also their work focuses on inter-table smells, whereas we focus on intra-table smells and their correction.

Harris et al. [16] implemented a tool that infers spreadsheet transformations by parsing a small example of the transformation and then extrapolating that example to an entire table. Their work focuses on transforming the layout of data cells and does not take cell formulas into account. REFBOOK has a predefined set of refactorings while their tool infers a new transformation for every example.

The inspiration for our project comes from research on refactoring for end-user programming in the context of Yahoo Pipes [29]. While both Yahoo Pipes refactoring and REFBOOK target end-users, the environments of these users are different. Spreadsheets have different smells and require a different set of tools to remove these smells.

VIII. CONCLUSIONS

End users working with spreadsheets make the same poor choices that professional developers make and have to pay the same “technical debt” that professional programmers pay during maintenance.

We designed, implemented, and evaluated REFBOOK, the first refactoring tool for spreadsheet formulas. It currently supports 7 refactorings that eliminate smells in spreadsheets.

Our three-pronged evaluation (case study of the EUSES Spreadsheet Corpus, user survey and controlled experiment with 28 participants) concludes that the refactorings supported by REFBOOK are widely applicable, increase programmer productivity, increase the reliability of transformations, and increase the quality of spreadsheets. More research is needed to find why end users do not feel comfortable with abstraction and how to create tools that they can embrace.

Acknowledgements: The authors thank Cosmin Radoi, Semih Okur, and the anonymous reviewers for constructive feedback, Jeff Overbey for providing assistance with Ludwig, Patrick O’Beirne for his advice on selecting refactorings to automate, the 28 participants in our study, and Microsoft for partially funding this research through a SEIF award.

REFERENCES

- [1] B. Boehm, C. Abts, A. Brown, S. Chulani, B. Clark, E. Horowitz, R. Madachy, J. Reifer, and Steece, *Software Cost Estimation with COCOMO II*. Prentice Hall PTR, 2000.
- [2] A. K. et al., “The state of the art in end-user software engineering,” *ACM Comput. Surv.*, vol. 43, no. 3, 2011.
- [3] C. Scaffidi, M. Shaw, and B. Myers, “Estimating the numbers of end users and end user programmers,” in *VLHCC*, 2005.
- [4] M. Burnett, C. Cook, and G. Rothermel, “End-user software engineering,” *Commun. ACM*, vol. 47, no. 9, 2004.
- [5] M. Fisher and G. Rothermel, “The euses spreadsheet corpus: a shared resource for supporting experimentation with spreadsheet dependability mechanisms,” *SIGSOFT Softw. Eng. Notes*, vol. 30, no. 4, 2005.
- [6] (2012, Jun.) Formuladatasleuth excel spreadsheet checking. <http://www.fairwayassociates.co.uk/formuladatasleuth>.
- [7] K. J. Rothermel, C. R. Cook, M. M. Burnett, J. Schonfeld, T. R. G. Green, and G. Rothermel, “Wysiwyt testing in the spreadsheet paradigm: an empirical evaluation,” in *ICSE*, 2000.
- [8] J. Cunha, J. P. Fernandes, J. Mendes, and J. S. Hugo Pacheco, “Towards a Catalog of Spreadsheet Smells,” in *ICCSA*, 2012.
- [9] F. Hermans, M. Pinzger, and A. van Deursen, “Supporting professional spreadsheet users by generating leveled dataflow diagrams,” in *ICSE*, 2011.
- [10] —, “Detecting and visualizing inter-worksheet smells in spreadsheets,” in *ICSE*, 2012.
- [11] F. Hermans, M. Pinzger, and A. V. Deursen, “Code smells in spreadsheet formulas,” in *ICSM*, 2012.
- [12] M. Erwig, “Software engineering for spreadsheets,” *IEEE Softw.*, vol. 26, no. 5, 2009.
- [13] J. F. Raffensperger, “New guidelines for spreadsheets,” *International Journal of Business and Economics*, vol. 2, no. 2, 2003.
- [14] P. O’Beirne, “Spreadsheet refactoring,” in *EuSpRiG*, 2010.
- [15] R. R. Panko, “What we know about spreadsheet errors,” *J. End User Comput.*, vol. 10, no. 2, 1998.
- [16] W. R. Harris and S. Gulwani, “Spreadsheet table transformations from examples,” in *PLDI*, 2011.
- [17] (2012, Jun.) Pup v7 utilities, the spreadsheet page. <http://spreadsheetpage.com/index.php/pupv7/utilities>.
- [18] (2012, Jun.) Asap utilities the essential add-in for excel. <http://www.asap-utilities.com/index.php>.
- [19] M. Fowler, *Refactoring: Improving the Design of Existing Code*. Addison-Wesley, 1999.
- [20] J. Ratzinger, T. Sigmund, and H. C. Gall, “On the relation of refactorings and software defect prediction,” in *MSR*, 2008.
- [21] (2012, Jun.) EuSpRiG: European Spreadsheet Risk Interest Group. <http://www.eusprig.org>.
- [22] (2012, Jun.) Openoffice.org forum. <http://www.ooffice.org/>.
- [23] (2012, Jun.) Excel help forum. <http://www.excelforum.org>.
- [24] J. L. Overbey and R. E. Johnson, “Generating rewritable abstract syntax trees,” in *SLE*, 2009.
- [25] S. Badame, “Refactoring meets spreadsheet formulas,” Master’s thesis, Univ of Illinois, <http://hdl.handle.net/2142/31155>, 2012.
- [26] (2012, Jun.) Apache openoffice calc suite page. <http://www.openoffice.org/product/calc.html>.
- [27] B. Opdyke, “Refactoring object-oriented frameworks,” Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1992.
- [28] (2012, Jun.) Apache poi. <http://poi.apache.org/>.
- [29] K. T. Stolee and S. Elbaum, “Refactoring pipe-like mashups for end-user programmers,” in *ICSE*, 2011.
- [30] M. Erwig and M. M. Burnett, “Adding apples and oranges,” in *PADL*, 2002.