SimilaR: R Code Clone and Plagiarism Detection

by Maciej Bartoszuk and Marek Gagolewski

Abstract
Third-party software for assuring source code quality is becoming increasingly popular. Tools that evaluate the coverage of unit tests, perform static code analysis, or inspect run-time memory use are crucial in the software development life cycle. More sophisticated methods allow for performing meta-analyses of large software repositories, e.g., to discover abstract topics they relate to or common design patterns applied by their developers. They may be useful in gaining a better understanding of the component interdependencies, avoiding cloned code as well as detecting plagiarism in programming classes.

A meaningful measure of similarity of computer programs often forms the basis of such tools. While there are a few noteworthy instruments for similarity assessment, none of them turns out particularly suitable for analysing R code chunks. Existing solutions rely on rather simple techniques and heuristics and fail to provide a user with the kind of sensitivity and specificity required for working with R scripts. In order to fill this gap, we propose a new algorithm based on a Program Dependence Graph, implemented in the SimilaR package. It can serve as a tool not only for improving R code quality but also for detecting plagiarism, even when it has been masked by applying some obfuscation techniques or imputing dead code. We demonstrate its accuracy and efficiency in a real-world case study.

Introduction
In recent years there has been a rise in the availability of tools related to code quality, including inspecting run-time memory usage (Serebryany et al., 2012), evaluating unit tests coverage (Ammann and Offutt, 2013), discovering abstract topics to which source code is related (Grant et al., 2012; Tian et al., 2009; McBurney et al., 2014; Linstead et al., 2007; Maskeri et al., 2008), finding parts of code related to a particular bug submission (Lukins et al., 2008), and checking for similarities between programs. With regards to the latter, quantitative measures of similarity between source code chunks play a key role in such practically important areas as software engineering, where encapsulating duplicated code fragments into functions or methods is considered a good development practice or in computing education, where any cases of plagiarism should be brought to a tutor’s attention, see (Misic et al., 2016; Mohd Noor et al., 2017; Roy et al., 2009; Rattan et al., 2013; Ali et al., 2011; Hage et al., 2011; Martins et al., 2014). Existing approaches towards code clone detection can be classified based on the abstraction level at which they inspect programs’ listings.

• Textual – the most straightforward representation, where a listing is taken as-is, i.e., as raw text. Typically, string distance metrics (like the Levenshtein one; see, e.g., van der Loo, 2014) are applied to measure similarity between pairs of the entities tested. Then some (possibly approximate) nearest neighbour search data structures seek matches within a larger code base. Hash functions can be used for the same purpose, where fingerprints of code fragments might make the comparison faster, see, e.g., (Johnson, 1993; Manber, 1994; Rieger, 2005). Another noteworthy approach involves the use of Latent Semantic Analysis (Marcus and Maletic, 2001) for finding natural clusters of code chunks.

• Lexical (token-based) – where a listing is transformed into tokens, which are generated by the parser during the lexical analysis stage. This form is believed to be more robust than the textual one, as it is invariant to particular coding styles (indentation, layout, comments, etc.). Typically, algorithms to detect and analyse common token sub-sequences are used (Kamiya et al., 2002; Ueda et al., 2002; Li et al., 2006; Wise, 1992; Prechelt et al., 2000; Hummel et al., 2011; Schleimer et al., 2003).

• Syntactic – an approach based on abstract syntax trees (ASTs). Contrary to the previous representation, which is “flat” by its nature, here the code is represented in a hierarchical manner. This makes it possible to, for instance, distinguish between a top-level statement and a code block that is executed within the else clause of an if-else construct. One way to quantify the similarity of ASTs is to find common sub-trees. This might be an uneasy task, therefore, e.g., in (Baxter et al., 1998) a hash function is used to project a sub-tree into a discrete single value, so that only the sub-trees in the same bucket are compared against each other. Another way involves the computation of diverse tree metrics (based on node type count, their distribution, connectivity, etc.) so that each AST is represented as a feature vector. Then the feature vectors
can be compared against each other directly in a pairwise fashion (Mayrand et al., 1996; Pate-naude et al., 1999; Fu et al., 2017) or by means of some cluster analysis-based approach (Jiang et al., 2007).

- **Semantic** – the most sophisticated representation involving a set of knowledge-based, language-dependent transformations of a program’s abstract syntax tree. Usually, a data structure commonly known as a Program Dependence Graph (PDG) is created, see below for more details. In such a data structure, the particular order of (control- or data-) independent code lines is negligible. A popular approach to measure similarity between a pair of PDGs concerns searching for (sub)isomorphisms of the graphs, see (Komondoor and Horwitz, 2001; Liu et al., 2006; Qu et al., 2014).

There are a few generally available software solutions whose purpose is to detect code clones, e.g., MOSS (see [http://theory.stanford.edu/~aiken/moss/](http://theory.stanford.edu/~aiken/moss/) and Schleimer et al., 2003) and JPlag (see [http://jplag.de/](http://jplag.de/) and Prechelt et al., 2000), see also (Misic et al., 2016; Vandana, 2018) for an overview. These tools are quite generic, offering built-in support for popular programming languages such as Java, C#, C++, C, or Python.

Unfortunately, there is no package of this kind that natively supports the R language, which is the GNU version of S (see, e.g., Becker et al., 1998; Venables and Ripley, 2000). It is a serious gap: R is amongst the most popular languages\(^1\), and its use has a long, successful track record, particularly with respect to all broadly-conceived statistical computing, machine learning, and other data science activities (Wickham and Grolemund, 2017). With some pre-processing, MOSS and JPlag can be applied on R code chunks, but the accuracy of code clones detection is far from optimal. This is due to the fact that, while at a first glance being an imperative language, R allows plenty typical functional constructs (see the next section for more details and also, e.g., Chambers, 1998; Wickham, 2014; Chambers, 2008). On the one hand, its syntax resembles that of the C language, with curly braces to denote a nested code block and classical control-flow expressions such as if...else conditionals, or while and for (for each) loops. On the other hand, R’s semantics is based on the functional Scheme language (Abelson et al., 1996), which is derived from Lisp. Every expression (even one involving the execution of a for loop) is in fact a call to a function or any combination thereof, and each function is a first-class object that (as a rule of thumb) has no side effects. Moreover, users might choose to prefer applying Map–Filter–Reduce-like expressions on container objects instead of the classical control-flow constructs or even mix the two approaches. Also, the possibility of performing the so-called nonstandard evaluation (metaprogramming) allows to change the meaning of certain expressions during run-time. For instance, the popular forward-pipe operator, \%\%\%, implemented in the magrittr (Bache and Wickham, 2014) package, allows for converting a pipeline of function calls to a mutually nested series of calls.

In this paper we describe a new algorithm that aims to fill the aforementioned gap (based on Bartoszuk, 2018). The method’s implementation is included in the SimilaR\(^2\) package. It transforms the analysed code base into a Program Dependence Graph that takes into account the most common R language features as well as the most popular development patterns in data science. Due to this, the algorithm is able to detect cases of plagiarism quite accurately. Moreover, thanks to a novel, polynomial-time approximate graph comparison algorithm, its implementation has relatively low run-times. This enables to conduct an analysis of a software repository whose size is significant.

This paper is set out as follows. First we introduce the concept of a Program Dependence Graph along with its R language-specific customisations. Then we depict a novel algorithm for quantifying similarity of two graphs. Further on we provide some illustrative examples for the purpose of showing the effects of applying particular alterations to a Program Dependence Graph. What is more, we demonstrate the main features of the SimilaR package version 1.0.8. Then we perform an experiment involving the comparison of the complete code-base of two CRAN packages.

**Program Dependence Graph**

A Program Dependence Graph (PDG) is a directed graph representing various relations between individual expressions in a source code chunk. As we mentioned in the introduction, it is among the most sophisticated data structures used for the purpose of code clones detection. First proposed by Ferrante et al. (1987), it forms the basis of many algorithms, see, e.g., (Liu et al., 2006; Qu et al., 2014; Gabel et al., 2008; Krinke, 2001; Horwitz and Reps, 1991; Komondoor and Horwitz, 2001; Ghosh and Lee, 2018; Nasirloo and Azimzadeh, 2018).


\(^2\)See [https://CRAN.R-project.org/package=SimilaR](https://CRAN.R-project.org/package=SimilaR). SimilaR can be downloaded from the Comprehensive R Archive Network (CRAN) repository ([Silge et al., 2018](http://jplag.de/)) and installed via a call to `install.packages("SimilaR")`. 
Abstract Syntax Tree. To create a PDG, we first need to construct an Abstract Syntax Tree (AST) of a given program. In R, it is particularly easy to compute the AST corresponding to any expression, due to its built-in support for reflection that facilitates metaprogramming. For instance, the `parse()` function can be called to perform lexical analysis of a code fragment, yielding a sequence of language objects. Moreover, a basic version of a function to print an AST takes just few lines of code:

```r
R> show_ast <- function(x) {
+     as.list_deep <- function(x) # convert to a plain list (recursively)
+     { if (is.call(x)) lapply(as.list(x), as.list_deep) else x }
+     x <- substitute(x) # expression that generated the argument
+     str(as.list_deep(x)) # pretty-print
+ }
```

Let us visualise the AST corresponding to expression `d <- sum((x-y)*(x-y))`.

```r
R> show_ast(d <- sum((x-y)*(x-y)))
```

List of 3
$ : symbol <-
$ : symbol d
$ : List of 2
  ..$ : symbol sum
  ..$ : List of 3
    ...$ : symbol *
    ....$ : List of 2
      ......$ : symbol -
      ......$ : symbol x
      ......$ : symbol y
    ....$ : List of 2
      ......$ : symbol -
      ......$ : symbol x
      ......$ : symbol y

In R, both a constant (numeric, logical, string, etc.) and a symbol (name) constitute what we call a *simple expression*. A *compound expression* is in turn a sequence of $n+1$ expressions (simple or compound ones) $\langle f, a_1, \ldots, a_n \rangle$, $n \geq 0$, which represents a call to $f$ with arguments $a_1, \ldots, a_n$ (which we are typically used to denote as $f(a_1, \ldots, a_n)$).

The above AST can be written in the Polish (prefix) notation as

```
\langle '<-', d, \langle 'sum', \langle '*', \langle '-' , x, y \rangle, \langle '-' , x, y \rangle \rangle \rangle \rangle
```

Such a notation is used in Scheme and Lisp; we skipped a call to `$()` for readability, as $(e)$ is equivalent to $e$ for each expression $e$. Alternatively, the above can be written as

```
'<-'(d, sum(\langle '*' , \langle '-' , x, y \rangle , \langle '-' , x, y \rangle \rangle \rangle)
```

in the “functional” form. Let us emphasise that even an application of a binary operator (here: `$<-`), `$*` and `$-` corresponds to some function call. Hence $x-y$ is just a syntactic sugar for $\langle '-' , x, y \rangle$. Moreover, other expressions such as if...else and loops also correspond to some function calls. For example:

```r
R> show_ast(for(i in 1:5) {
+     print("i = ", i)
+     if (i %% 2 == 0) print(":)") else print(":$")
+ })
```

List of 4
$ : symbol for
$ : symbol i
$ : List of 3
  ..$ : symbol :
  ..$ : num 1
  ..$ : num 5
Vertex and edge types. The vertices of a PDG represent particular expressions, such as a variable assignment, a function call or a loop header. Each vertex is assigned its own type, reflecting the kind of expression it represents. The comprehensive list of vertex types for the R language code-base used in SimilaR is given in Table 1. The number of distinct types is a kind of compromise between the algorithm’s sensitivity and specificity. It was set empirically based on numerous experiments (Bartoszuk, 2018).

We may also distinguish two types of edges: control dependency and data dependency ones. The former represents the branches in a program’s control flow that result in a conditional execution of expressions such as if-else-constructs or loops. A subgraph of a PDG consisting of all the vertices and only the control dependency edges is called a Control Dependence Subgraph (CDS).

The latter edge vertex type is responsible for modelling data flow relations: there is an edge from a vertex \( v \) to a vertex \( u \), whenever a variable assigned in the expression corresponding to \( v \) is used in the expression corresponding to \( u \).

<table>
<thead>
<tr>
<th>Id</th>
<th>Color</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Olive</td>
<td>Entry</td>
<td>marks the beginning of a function</td>
</tr>
<tr>
<td>1</td>
<td>Light yellow</td>
<td>Header</td>
<td>loop</td>
</tr>
<tr>
<td>2</td>
<td>–</td>
<td>Next</td>
<td>next</td>
</tr>
<tr>
<td>3</td>
<td>–</td>
<td>Break</td>
<td>break</td>
</tr>
<tr>
<td>4</td>
<td>Orange</td>
<td>If</td>
<td>a conditional expression</td>
</tr>
<tr>
<td>5</td>
<td>Light blue</td>
<td>If_part</td>
<td>an expression to execute conditionally</td>
</tr>
<tr>
<td>6</td>
<td>Gray</td>
<td>assignment</td>
<td>an assignment expression, e.g., name &lt;- val</td>
</tr>
<tr>
<td>7</td>
<td>Violet</td>
<td>parameter</td>
<td>a function parameter</td>
</tr>
<tr>
<td>8</td>
<td>–</td>
<td>oneBracketSingle</td>
<td>an expression involving ([i]), e.g., vector([-1])</td>
</tr>
<tr>
<td>9</td>
<td>–</td>
<td>oneBracketDouble</td>
<td>an expression involving ([i,j])</td>
</tr>
<tr>
<td>10</td>
<td>–</td>
<td>oneBracketTripleOrMore</td>
<td>an expression involving ([i,j,k,\ldots])</td>
</tr>
<tr>
<td>11</td>
<td>–</td>
<td>twoBrackets</td>
<td>an expression involving ({1}), e.g., list[&quot;nameditem&quot;]]</td>
</tr>
<tr>
<td>12</td>
<td>–</td>
<td>dollar</td>
<td>an expression involving $, e.g., df$column</td>
</tr>
<tr>
<td>13</td>
<td>–</td>
<td>funNoArguments</td>
<td>a function call with no actual parameters</td>
</tr>
<tr>
<td>14</td>
<td>Red</td>
<td>FunOneArgument</td>
<td>a function call with one actual parameter</td>
</tr>
<tr>
<td>15</td>
<td>–</td>
<td>FunTwoArguments</td>
<td>a function call with two actual parameters</td>
</tr>
<tr>
<td>16</td>
<td>–</td>
<td>FunThreeArguments</td>
<td>a function call with three actual parameters</td>
</tr>
<tr>
<td>17</td>
<td>–</td>
<td>FunFourOrMoreArguments</td>
<td>a function call with four or more actual parameters</td>
</tr>
<tr>
<td>18</td>
<td>–</td>
<td>stopifnot</td>
<td>a call to stopifnot()</td>
</tr>
<tr>
<td>19</td>
<td>Green</td>
<td>arithmeticOperator</td>
<td>an arithmetic operator-based expression, e.g., +, -, *</td>
</tr>
<tr>
<td>20</td>
<td>Blue</td>
<td>comparisonOperator</td>
<td>a comparison operator-based expression, e.g., ==, &lt;= or &lt;</td>
</tr>
<tr>
<td>21</td>
<td>Green</td>
<td>return</td>
<td>a call to return()</td>
</tr>
<tr>
<td>22</td>
<td>Cyan</td>
<td>colon</td>
<td>a colon operator-based expression, e.g., from:to</td>
</tr>
<tr>
<td>23</td>
<td>Dark green</td>
<td>symbol</td>
<td>a symbol (name)</td>
</tr>
<tr>
<td>24</td>
<td>–</td>
<td>constant</td>
<td>a constant value, e.g., 1, &quot;string&quot; or NA</td>
</tr>
</tbody>
</table>

Table 1: Assumed Program Dependence Graph vertex types for the R language.
computation of the expression related to \( u \). A spanning subgraph of a PDG that consists solely of the data dependency edges is called a Data Dependence Subgraph (DDS).

Hence, a PDG of a function \( F() \) is a vertex- and edge-labelled directed graph \( F = (V_F, E_F, \zeta_F, \xi_F) \), where \( V_F \) is the set of its vertices, \( E_F \subseteq V_F \times V_F \) denotes the set of edges \((v, u) \in E_F \) whenever there is an edge from \( v \) to \( u \), \( \zeta_F : V_F \to \{\text{Entry}, \text{Header}, \ldots, \text{constant}\} = \{0, \ldots, 25\} \) gives the type of each vertex and \( \xi_F : E_F \to \{\text{DATA}, \text{CONTROL}\} \) marks if an edge is a data- or control-dependency one. Note that each PDG is rooted – there exists one and only one vertex \( v \) with indegree 0 and \( \zeta_F(v) = \text{Entry} \).

Example code chunks with the corresponding dependence graphs are depicted in Figures 1 and 2. The meaning of vertex colors is explained in Table 1.

\[
\text{sum} <- \text{function}(x) \\
\{ \\
\quad s <- 0 \\
\quad m <- 1 \\
\quad \text{for}(i \in x) \{ \\
\quad \quad s <- s + i \\
\quad \quad m <- m \times i \\
\quad \} \\
\quad \text{if}(s < 0) \{ \\
\quad \quad s <- -s \\
\quad \quad \text{print("Negative s")} \\
\quad \} \\
\quad \text{if}(m < 0) \{ \\
\quad \quad m <- -m \\
\quad \quad \text{print("Negative m")} \\
\quad \} \\
\quad \text{return}(s) \\
\}
\]

**Figure 1:** An example function and the respective Control Dependence Subgraph.

\[
\text{sum} <- \text{function}(x) \\
\{ \\
\quad a <- 5 \\
\quad b <- 6 \\
\quad \text{for}(i \in x) \{ \\
\quad \quad c <- a + b - i \\
\quad \} \\
\}
\]

**Figure 2:** An example function and the corresponding Program Dependence Graph; solid and dashed arrows represent control and data dependency edges, respectively.
The most basic version of an algorithm to create a PDG based on an abstract syntax tree is described in (Harrold et al., 1993). Let us note the fact that a CDS is a subgraph of an AST: it provides the information about certain expressions being nested within other ones, e.g., that some assignment is part (child) of a loop’s body. Additionally, an AST includes a list of local variables and links them with expressions that rely on them. This is a crucial piece of information used to generate DDS.

Note that, however, a PDG created for the purpose of code clones detection cannot be treated as a straightforward extension of a raw AST. The post-processing procedure should be carefully customised taking into account the design patterns and coding practices of a particular programming language. Hence, below we describe the most noteworthy program transforms employed in the SimilaR package so that it is invariant to typical attacks, i.e., transforms changing the way the code is written yet not affecting its meaning.

### Unwinding nested function calls.

As mentioned above, in R, as in any functional language, functions play a key role. A code chunk can be thought of as a sequence of expressions, each of which is composed of function calls. Base or external library functions are used as a program’s building blocks and often very complex tasks can be written with only few lines of code.

For instance, given a matrix $X \in \mathbb{R}^{d \times n}$ representing $n$ vectors in $\mathbb{R}^d$ and a vector $y \in \mathbb{R}^d$, the closest vector in $X$ to $y$ with respect to the Euclidean metric can be determined by evaluating $X[, \text{which.min(apply((X-y)^2,2,sum))}]$. This notation is very concise and we can come up with many equivalent forms of this expression written in a much more loquacious fashion.

Therefore, in SimilaR, hierarchies of nested calls, no matter their depth, are always recursively unwound by introducing as many auxiliary assignments as necessary. For instance, $f(g(x))$ is decomposed as $gx \leftarrow g(x); f(gx)$. This guarantees that all their possible variants are represented in the same way in the PDG.

### Forward-pipe operator, $%>%$

Related to the above is the magrittr’s forward-pipe operator, $%>%$, which has recently gained much popularity within the R users’ community. Even though the operator is just a syntactic sugar for forwarding an object into the next function call/expression, at the time of writing of this manuscript, the package has been used as a direct dependency by over 700 other CRAN packages. Many consider it very convenient, as it mimics the “left-to-right” approach known from object-orientated languages like Java, Python or C++. Instead of writing (from the inside and out) $f(g(x),y)$, with magrittr we can use the syntax $x %>% g %>% f(y)$ (which would normally be represented as $x.g().f(y)$ in other languages). To assure proper similarity evaluation, SimilaR unwinds such expressions in the same manner as nested function calls.

### Calls within conditions in control-flow expressions.

An expression serving as a Boolean condition in an if or a while construct might be given as a composition of many function calls. The same might be true for an expression generating the container to iterate over in a for loop. PDG vertices representing such calls are placed on the same level as their corresponding control-flow expressions so that they can be unwound just as any other function call.

### Canonicalization of conditional statements.

The following code chunk:

```r
if (cond) {
  return(A)
} else {
  return(B)
}
```

is semantically equivalent to:

```r
if (cond)
  return(A)
else
  return(B)
```

This exploits the fact that the code after the if statement is only executed whenever the logical condition is false. To avoid generating very different control dependencies, we always unwind the former to the latter by putting the code-richer branch outside of a conditional statement.
Tail call to `return()`. The return value that is generated by evaluating a sequence of expressions wrapped inside curly braces (the `{` function) is determined by the value of its last expression. If a function’s body is comprised of such a code block, the call to `return()` is optional if it is used in the last expression. However, many users write it anyway. Therefore, a special vertex of type `return()` have been introduced to mark an expression that generates the output of a function.

**Map-like functions.** Base R supports numerous Map-like operations that are available in many programming languages. The aim of the members of the `*apply()` family (`apply()`, `lapply()`, `sapply()`, etc.) is to perform a given operation on each element/slice of a given container. These are unwound as an expression involving a for loop. For instance, a call to `ret <- lapply(l, fun, ...)` can be written as `ret <- list(); for (el in l) ret[[length(ret)+1]] <- fun(el, ...)`. 

**Variable duplication.** To prevent redundant assignments such as `xcopy <- x` made just in order to refer to the original value under a new alias, a hierarchical variable dictionary is kept to generate data dependency edges properly.

**Memoization.** In pure functional languages it is assumed that functions have no side effects, i.e., the same arguments are mapped to the same return value. In R it is of course not always technically true (e.g., when the pseudo-random number generator is involved), but such an assumption turns out to be helpful in our context. Therefore, if a function call instance is invoked more than once, its value is memorised by introducing a new variable.

**Dead code.** Many plagiarism detection algorithms can be easily misled by adding random code that does not affect the main computations. In SimilaR, such dead code is identified and removed. This is done by iteratively deleting all vertices whose outdegree is zero (except those of type `return`).

To sum up, SimilaR guarantees that the Program Dependence Graph is the same regardless of the order of independent function calls, unwinding nested function calls, the use of the forward-pipe operator, etc. Hence, it is invariant to the most typical attacks. Moreover, it has been implemented in such a way that new kinds of transformations can be easily added in the future, as the R development practices and common program design patterns evolve.

### Comparing Program Dependence Graphs

In our setting, code similarity assessment reduces to a comparison between a pair of Program Dependence Graphs. In this section we are interested in an algorithm $\mu$ such that $\mu(F, G) \in [0, 1]$ represents a similarity degree between two PDGs $F$ and $G$. A similarity of 1 denotes that two PDGs are identical, while 0 means that they are totally different. Alternatively, we might be interested in a non-symmetric measure $\tilde{\mu}(F, G) \in [0, 1]$ representing the degree to which the source code of $F$ is contained within $G$.

Ideally, an algorithm to compare two PDGs should enjoy the following properties:

- it should be *flexible* in the sense that introducing a “small difference” in one of the graphs should not affect the estimated similarity degree significantly;
- it should be *fast* to execute so that computing numerous pairwise similarities can be performed in a reasonable time span.

Due to the latter, we immediately lose our interest in all currently known exact algorithms to find subgraph isomorphisms or maximum common subgraphs because of their exponential-time complexity (the problems are NP-hard; see, e.g., Wegener, 2005). To recall, two graphs are isomorphic whenever there exists a mapping between the two graphs’ vertices preserving the node adjacencies.

In the SimilaR package, we use a modified version (for increased flexibility and better performance in the plagiarism detection problem) of an algorithm described in (Shervashidze et al., 2011), which itself is based on the Weisfeiler–Lehman isomorphism test (Weisfeiler and Lehman, 1968) and graphs kernels. Note that the base method has been successfully used in many applications, e.g., in cheminformatics (Mapar, 2018) and programming autonomous robots (Luperto and Amigoni, 2019).

In each of the $h$ iterations of the SimilaR algorithm, we assign new labels to the vertices of a PDG based on their neighbours’ labels. While in the original algorithm (Shervashidze et al., 2011), two vertices are considered diverse already when one of their neighbours has been assigned a different...
label, here we might still be assigning the same label if the vertices’ adjacency differs only slightly. Our approach turns out to be more robust (Bartoszuk, 2018) against minor code changes or some vertices being missing in the graph.

**SimilaR algorithm at a glance.** Before we describe every step of the algorithm in detail, let us take a look at it from a bird’s-eye perspective. If we are to assign each vertex a new label that is uniquely determined by their current type as well the labels allocated to their neighbours, two identical graphs will always be coloured the same way, no matter how many times we reiterate the labelling procedure. In particular, after \( h \) iterations, a vertex’s label depends on the types of vertices whose distance from it is at most \( h \).

We are of course interested in assigning equivalent labels to vertices in graphs that are not necessarily identical, but still similar to each other. Otherwise, two vertices which have all but one neighbour in common, would get distinct labels. After \( h \) iterations, all the vertices at distance at most \( h \) would all already be assigned different colours. This would negatively affect the overall graph similarity assessment.

In order to overcome this problem, we introduce the concept of vertex importance, which is based upon the number of vertices that depend on a given node. Only important enough differences in the vertex neighbourhoods will be considered as sufficient to trigger a different labelling. Then, after \( h \) iterations, two vectors of label type counts can be compared with each other to arrive at a final graph similarity degree.

**SimilaR algorithm in detail.** The following description of the SimilaR algorithm will be illustrated based on a comparison between two functions, \texttt{clamp1()} (whose PDG from now on we will denote with \( F = (V_F, E_F, \zeta_F, \xi_F) \), see Fig. 3) and \texttt{standardise()} (denoted \( G = (V_G, E_G, \zeta_G, \xi_G) \), see Fig. 4).

<table>
<thead>
<tr>
<th>( v )</th>
<th>Entry</th>
<th>( x )</th>
<th>( \min(x) )</th>
<th>( \max(x) )</th>
<th>- (1)</th>
<th>( \text{If} )</th>
<th>( \text{If}_\text{part} )</th>
<th>NULL</th>
<th>- (2)</th>
<th>/</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw ( \delta_F(v) )</td>
<td>7.80</td>
<td>3.04</td>
<td>1.35</td>
<td>1.12</td>
<td>0.93</td>
<td>0.65</td>
<td>0.30</td>
<td>0.20</td>
<td>0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>normalised</td>
<td>0.49</td>
<td>0.19</td>
<td>0.09</td>
<td>0.07</td>
<td>0.06</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Table 2:** Vertex importance degrees in \texttt{clamp1()} (sum = 15.79, median = 0.04).

| \( v \) | Entry | \( x \) | \( \text{sd}(x) \) | \( < \) | \( \text{If} \) | \( \text{If}_\text{part} \) | NULL | mean(x) | - | / |
|-------|-------|-------|---------------|-------|-------|-------------|-------|--------|-------|
| raw \( \delta_G(v) \) | 4.33 | 1.71 | 0.93 | 0.65 | 0.30 | 0.20 | 0.10 | 0.33 | 0.21 | 0.1 |
| normalised | 0.49 | 0.19 | 0.10 | 0.07 | 0.03 | 0.02 | 0.01 | 0.04 | 0.02 | 0.01 |

**Table 3:** Vertex importance degrees in \texttt{standardise()} (sum = 8.86, median = 0.03).

| \( v \) | Entry | \( x \) | \( \min(x) \) | \( \max(x) \) | - (1) | \( \text{If} \) | \( \text{If}_\text{part} \) | NULL | mean(x) | - | / |
|-------|-------|-------|---------------|---------------|-------|-------|-------------|-------|--------|-------|
| \( \zeta_F(v) \) | 0 | 24 | 14 | 14 | 20 | 21 | 4 | 5 | 25 | 20 | 20 |
| \( \zeta_1^2(v) \) | 0 | 1 | 2 | 2 | 3 | 4 | 5 | 6 | 7 | 3 | 3 |
| \( \zeta_2^2(v) \) | 7 | 1 | 2 | 2 | 8 | 3 | 4 | 5 | 0 | 6 | 6 |
| \( \zeta_3^2(v) \) | 8 | 5 | 6 | 6 | 7 | 1 | 2 | 3 | 0 | 4 | 4 |

**Table 4:** \texttt{clamp1()}: Labels assigned to vertices in each iteration.

| \( v \) | Entry | \( x \) | \( \text{sd}(x) \) | \( < \) | \( \text{If} \) | \( \text{If}_\text{part} \) | NULL | mean(x) | - | / |
|-------|-------|-------|---------------|-------|-------|-------------|-------|--------|-------|
| \( \zeta_G(v) \) | 0 | 24 | 14 | 21 | 4 | 5 | 25 | 14 | 20 | 20 |
| \( \zeta_1^3(v) \) | 0 | 1 | 2 | 4 | 5 | 6 | 7 | 2 | 3 | 3 |
| \( \zeta_2^3(v) \) | 7 | 1 | 9 | 3 | 4 | 5 | 0 | 2 | 6 | 6 |
| \( \zeta_3^3(v) \) | 12 | 9 | 10 | 1 | 2 | 3 | 0 | 11 | 4 | 4 |

**Table 5:** \texttt{standardise()}: Labels assigned to vertices in each iteration.
CONTRIBUTED RESEARCH ARTICLE

\[ R> \text{clamp1} <- \text{function}(x) \]
\[ + \{ \]
\[ + \quad \text{if} (\text{max}(x) - \text{min}(x) < 1e-5) \]
\[ + \quad \text{NULL} \]
\[ + \quad \text{else} \]
\[ + \quad (x - \text{min}(x))/(\text{max}(x) - \text{min}(x)) \]
\[ + \} \]

\[ R> \text{standardise} <- \text{function}(x) \]
\[ + \{ \]
\[ + \quad \text{if} (\text{sd}(x) < 1e-5) \]
\[ + \quad \text{NULL} \]
\[ + \quad \text{else} \]
\[ + \quad (x - \text{mean}(x))/\text{sd}(x) \]
\[ + \} \]

Figure 3: The `clamp1()` function with the respective PDG.

Figure 4: The `standardise()` function with the respective PDG.

1. **Vertex importance degrees.** Firstly, each vertex in \( F \) is assigned an importance degree, \( \delta_F : V_F \to \mathbb{R}_+ \),
\[
\delta_F(v) = 0.1 + \sum_{(v,u) \in E_F} \left( 1 + 0.1 \text{I}_{\text{cond}}((v,u)) = \text{DATA} \right) \delta_F(u),
\]
where \( \text{I}_{\text{cond}} \) is an indicator function with value 1 if \( \text{cond} \) is true and 0 otherwise. In other words, a vertex \( v \) with outdegree equal to 0 has importance \( \delta(v) = 0.1 \). Otherwise, its importance degree is set to \( \delta(v) = 0.1 \) plus the sum of importances of its outgoing control-dependent neighbours plus the sum of importances of its outgoing data-dependent neighbours multiplied by 1.1.

Note that if \( F \) is an acyclic graph, then it has a topological ordering, i.e., an arrangement of the vertices such that every edge is directed from earlier to later in the sequence. In such a case \( \delta_F \) is well-defined. Otherwise, we shall be computing the importance degrees in the depth-first manner.

Next, the importance degrees are normalised, $\delta_F(v) \rightarrow \delta_F(v)/\sum_{u \in V_F} \delta_F(u)$. Tables 2 and 3 give the importance degrees of the vertices in the two graphs studied.

2. Vertex labels. Recall from the previous section that each vertex $v \in V_F, u \in V_G$ has been assigned a label, $\xi_F(v), \xi_G(u) \in \{0, \ldots, 25\}$, based on the type of operation it represents (see Tables 4 and 5).

In the $i$-th (out of $h$ in total, here we fix $h = 3$; see Bartoszuk, 2018 for discussion) iteration of the SimiRaR algorithm, we assign new labels $\xi_F^i, \xi_G^i$ according to the labels previously considered.

(a) Iteration $i = 1$. In the first iteration, the initial labels, $\xi_F, \xi_G$, are simply remapped to consecutive integers. This yields $\xi_F^1$ and $\xi_G^1$ as given in Tables 4 and 5.

(b) Iterations $i = 2$ and $i = 3$. In subsequent iterations, we seek groups of similar vertices so as to assign them the same label. Two vertices $v, u \in V_F \cup V_G$ are considered similar (with no loss in generality, we are assuming $v \in V_F$ and $u \in V_G$ below), whenever they have been assigned the same label in the previous iteration and have outgoing neighbours with the same labels. However, for greater flexibility, we allow for the neighbourhoods to differ slightly – unmatched neighbours of lesser importance degrees will not be considered significant. Formally, $v \in V_F$ and $u \in V_G$ are similar, whenever $\xi_F^{-1}(v) = \xi_G^{-1}(u)$ and:

$$
\sum_{(v,w) \in E_F, \xi_F^{-1}(w) \notin C(v,w)} \delta_F(w) \leq \min \{M_F, M_G\},
$$

where $C(v,u) = \{\xi_F^{-1}(w) : (v,w) \in E_F \} \cap \{\xi_G^{-1}(w) : (u,w) \in E_F\}$ denotes the multiset of common neighbours’ vertex labels and $M_F$ and $M_G$ denote the medians of importance degrees of vertices in $F$ and $G$, respectively.

The above similarity relation in obviously reflexive and symmetric. When we compute its transitive closure, we get an equivalence relation whose equivalence classes determine sets of vertices that shall obtain identical labels.

For instance, let $i = 2$ and $v$ be the Entry vertex in $F$ and $u$ be the Entry vertex in $u$. We have $\xi_F^1(v) = \xi_G^1(u) = 0$. Outgoing neighbours of $v$ have types: 1 $(x)$, 3 $(2)$, 3 $(1)$, 2 $(\min(x))$, 2 $(\max(x))$, 4 $(<)$, 3 $(\sim)$ and 5 $(\leq)$. Neighbours of $u$ are labelled as 1 $(x)$, 3 $(\sim)$, 2 $(\min(x))$, 3 $(\leq)$, 2 $(\max(x))$, 4 $(<)$, 5 $(\leq)$. Hence, $v$ has one unmatched neighbour of type 3 (here: an arithmetic operation). However, as there are 3 such neighbours of $v$ (importances: 0.01, 0.01, 0.06) and 2 of $u$ (importances: 0.01, 0.02), we need a rule for determining the left-out importance degree. Here, for greater algorithm’s flexibility, we always assume that the unmatched importances are considered in an increasing order. Therefore, we get that $0.01 \leq \min \{M_F, M_G\} = \min \{0.04, 0.03\} = 0.03$ and thus $v$ and $u$ are considered similar.

As another example, let $i = 3$ and $v$ be the $x$ vertex in $F$ and $u$ be the $x$ vertex in $G$. We have $\xi_F^2(v) = \xi_G^2(u) = 1$. Their neighbours are: $- 2$ (importance=0.01; type=6), $\min(x)$ (importance=0.09; type=2), $\max(x)$ (importance=0.07; type=2) and $- 2$ (importance=0.02; type=6), $\max(x)$ (importance=0.04; type=2), $\min(x)$ (importance=0.10; type=9). The sum of the importance degrees of the unmatched neighbours is now equal to 0.07 + 0.10 > 0.03, hence, $u$ and $v$ are not considered similar. They are also not in the transitive closure with respect to the other similarities. Therefore, they will be assigned different labels.

3. Partial similarity degrees. Let $m$ be the maximal integer label assigned above and $L_F = (L_{F,1}, \ldots, L_{F,m})$ be a vector of label counts, where $L_{F,j} = |\{v \in V_F : \xi_F^i(v) = j\}|$, see Table 6. We define $L_G$ in much the same way, see Table 7. We introduce the following “partial” similarity measures of the label sequences – the symmetric:

$$
\mu(L_F, L_G) = 1 - \frac{\sum_{k=1}^m |L_{F,k} - L_{G,k}|}{\sum_{k=1}^m L_{F,k} + \sum_{k=1}^m L_{G,k}}
$$

and its nonsymmetric version:

$$
\bar{\mu}(L_F, L_G) = \frac{\sum_{k=1}^m \min(L_{F,k}, L_{G,k})}{\sum_{k=1}^m L_{F,k}}.
$$

The partial similarities for $i = 1, 2, 3$ are given in Table 8.

4. Final similarity degrees. The overall similarity degree is defined as the arithmetic mean of the $h = 3$ partial similarities (reported in Table 8):

$$
\mu(F, G) = \frac{1}{h} \sum_{i=1}^h \mu(L_F, L_G) = \frac{0.95 + 0.86 + 0.57}{3} = 0.79,
$$

(1)
Table 6: clamp1(): Label counts in each iteration.

<table>
<thead>
<tr>
<th>j</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{F,j}^1$</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$L_{F,j}^2$</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$L_{F,j}^3$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: standardise(): Label counts in each iteration.

<table>
<thead>
<tr>
<th>j</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{G,j}^1$</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$L_{G,j}^2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$L_{G,j}^3$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Similarity measures in each iteration.

<table>
<thead>
<tr>
<th>i</th>
<th>$\mu(L_{F,j}^i, L_{G,j}^i)$</th>
<th>$\hat{\mu}(L_{F,j}^i, L_{G,j}^i)$</th>
<th>$\hat{\mu}(L_{G,j}^i, L_{F,j}^i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.95</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0.86</td>
<td>0.82</td>
<td>0.90</td>
</tr>
<tr>
<td>3</td>
<td>0.57</td>
<td>0.55</td>
<td>0.60</td>
</tr>
<tr>
<td>Final</td>
<td>0.79</td>
<td>0.76</td>
<td>0.83</td>
</tr>
</tbody>
</table>

We obtain its nonsymmetric versions in the same way:

$$\hat{\mu}(F, G) = \frac{1}{h} \sum_{i=1}^{h} \hat{\mu}(L_{F,j}^i, L_{G,j}^i) = \frac{0.91 + 0.82 + 0.55}{3} = 0.76,$$

$$\hat{\mu}(G, F) = \frac{1}{h} \sum_{i=1}^{h} \hat{\mu}(L_{G,j}^i, L_{F,j}^i) = \frac{1.00 + 0.90 + 0.60}{3} = 0.83.$$

(2)

Illustrative examples

The SimilaR package can be downloaded from CRAN and installed on the local system via a call to:

```R
R> install.packages("SimilaR")
```

Here we are working with version 1.0.8 of the package.

Once the package is loaded and its namespace is attached by calling:

```R
R> library("SimilaR")
```

two functions are made available to a user. SimilaR_fromTwoFunctions() is responsible for assessing the similarity between a pair of function objects (R is a functional language, hence assuming that functions constitute basic units of code seem natural). Moreover, SimilaR_fromDirectory(), which we shall use in the next section, is a conveniently vectorised version of the former, performing the comparison of all the scripts in a given directory:

Let us evaluate the similarity between the $F = \text{clamp1}()$ and $G = \text{standardise}()$ functions defined above:

```R
R> SimilaR_fromTwoFunctions(clamp1, standardise) # aggregation="sym"

name1 name2 SimilaR decision
clamp1 standardise 0.7954545 1
```

The R Journal Vol. XX/YY, AAAA 20ZZ ISSN 2073-4859
Here, \( \text{SimilaR} \) denotes the (symmetric) measure \( \mu(F, G) \in [0, 1] \) as given by Eq. (1). decision uses a built-in classifier model to assess whether such a similarity degree is considered significant (1) or not (0).

To obtain the non-symmetric measures, \( \tilde{\mu}(F, G) \) and \( \tilde{\mu}(G, F) \) (see Eq. (2)), we pass \texttt{aggregation="both"} as an argument:

```r
R> SimilaR_fromTwoFunctions(clamp1, standardise, aggregation="both")
```

<table>
<thead>
<tr>
<th>name1</th>
<th>name2</th>
<th>SimilaR12</th>
<th>SimilaR21</th>
<th>decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>clamp1</td>
<td>standardise</td>
<td>0.7575758</td>
<td>0.8333333</td>
<td>1</td>
</tr>
</tbody>
</table>

**Example:** \texttt{clamp2()}. Let us recall the source code of the \texttt{clamp1()} function:

```r
R> clamp1
function(x)
{
  if (max(x)-min(x) < 1e-5)
    NULL
  else
    (x - min(x))/(max(x)-min(x))
}
```

By applying numerous transformations described above, we may arrive at its following version:

```r
R> library("magrittr")
R> clamp2 <- function(y)
+ {
+   longName <- y # variable duplication
+   longName2 <- min
+   z <- { sum(longName**2) } # dead code
+   min_y <- longName %>% longName2 # forward-pipe
+   max_y <- y %>% max
+   max_y_min_y <- max_y-min_y # memoization
+   if(!(max_y_min_y >= 1e-5)) # canonicalization of the if statement
+   {
+     return(NULL)
+   }
+   ((y - min_y)/max_y_min_y) # tail call to return removed
+ }
```

\texttt{SimilaR} correctly identifies the two functions as equivalent:

```r
R> SimilaR_fromTwoFunctions(clamp1, clamp2, aggregation="both")
```

<table>
<thead>
<tr>
<th>name1</th>
<th>name2</th>
<th>SimilaR12</th>
<th>SimilaR21</th>
<th>decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>clamp1</td>
<td>clamp2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Example:** A vectorised version of \texttt{clamp1()}. Let us now consider two different vectorised versions of the \texttt{clamp1()} function for list-type inputs. The first one is based on a call to \texttt{lapply()}, which takes care of applying a given anonymous function on each list member:

```r
R> clamp1_vectorised1 <- function(x) {
+   x %>% lapply(function(y) {
+     if (max(y)-min(y) < 1e-5) {
+       {{{{NULL}}}}
+     } else {
+       {{{{(y - min(y))/(max(y)-min(y))}}}}
+     }
+   })
+ }
```

The second function unwinds the Map-like construct, adding a \texttt{for}-loop instead. It also relies on some domain knowledge, namely, that a new list is pre-allocated with \texttt{NULL}s.
CONTRIBUTED RESEARCH ARTICLE

```r
R> clamp1_vectorised2 <- function(x) {
+ n <- length(x)
+ res <- vector("list", n) # NULLs
+ for (i in 1:n) { # assumed n>0
+ m <- min(x[[i]])
+ mm <- max(x[[i]])-m
+ if (mm >= 1e-5)
+ res[[i]] <- (x[[i]] - m)/mm
+ }
+ return(res)
+ }

Note that the function only works for non-empty input lists.

The pairwise comparison yields:

```r
R> SimilaR_fromTwoFunctions(clamp1_vectorised1, clamp1_vectorised2,
+ aggregation="both")
```

<table>
<thead>
<tr>
<th>name1</th>
<th>name2</th>
<th>SimilaR12</th>
<th>SimilaR21</th>
<th>decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>clamp1_vectorised1</td>
<td>clamp1_vectorised2</td>
<td>0.7833333</td>
<td>0.9215686</td>
<td>1</td>
</tr>
</tbody>
</table>

Which indicates a significant degree of similarity, which is indeed the case.

A case study

In the previous section we illustrated that SimilaR is easily able to identify the code chunks that can be transformed onto each other. Now we shall demonstrate its usefulness in a real-world scenario: let us compare the code-base of two R packages: nortest (Gross and Ligges, 2015) and DescTools (Signorell et al., 2020). The former implements five significance tests for normality, while the latter advertises itself as

A collection of miscellaneous basic statistic functions and convenience wrappers for efficiently describing data. […] Many of the included functions can be found scattered in other packages and other sources written partly by Titans of R. The reason for collecting them here, was primarily to have them consolidated in ONE instead of dozens of packages (which themselves might depend on other packages which are not needed at all), and to provide a common and consistent interface as far as function and arguments naming, NA handling, recycling rules, etc. are concerned. […] (DescTools package DESCRIPTION; Signorell et al., 2020)

1. Set-up. First we attach the required packages and set up the directory where we shall store the data that we are going to feed the algorithm with at a later stage.

```r
R> library("SimilaR")
R> dir_output <- tempfile("dir")
R> dir.create(dir_output)
```

2. Generate code-base. SimilaR does not offer direct support for comparing two R packages. Therefore, below we export each package’s code-base to a single source file, getting rid of non-function objects:

```r
R> for (pkg in c("DescTools", "nortest"))
+ {
+ library(pkg, character.only=TRUE) # attach the package
+ env <- as.environment(paste0("package:", pkg)) # package's environment
+ fun_names <- ls(envir=env) # list of all exported objects
+ file_text <- character(0) # the to-be code-base (1 function == 1 string)
+ for (fun in fun_names)
+ {
+ f <- get(fun, env) # get function object
+ if (is.function(f)) next
+ f_char <- paste0(deparse(f), collapse="\n") # extract source code
+ ```
+ file_text[length(file_text)+1] <- sprintf("%s\n", pkg, length(file_text)))
+ }

DescTools: 549 functions processed.
nortest: 5 functions processed.

Here the list of the objects exported by both packages is determined by querying their corresponding package:DescTools and package:nortest environments. Moreover, a call to deparse() on a function object gives a plain-text representation of its source code.

3. Run the algorithm. Now we ask the algorithm to fetch the two source files in the output directory and execute all the pairwise comparisons between the functions defined therein.

R> time0 <- Sys.time()
R> results <- SimilaR_fromDirectory(dir_output, + fileTypes="file", aggregation="both")
R> print(Sys.time()-time0)

Time difference of 15.41459 secs

The above gives the time to execute all the 2745 pairwise comparisons on an Intel Core i7 laptop with 16GB RAM, running the GNU/Linux 4.19.30-041930-generic SMP x86_64 kernel.

4. Report results. Let us inspect the top 10 results returned by the algorithm (in terms of overall similarity).

R> print(head(results, 10))

For greater readability, the results are reported in Table 9.

<table>
<thead>
<tr>
<th>Name1 (DescTools)</th>
<th>Name2 (nortest)</th>
<th>SimilaR12</th>
<th>SimilaR21</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>CramerVonMisesTest()</td>
<td>cvm.test()</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1</td>
</tr>
<tr>
<td>LillieTest()</td>
<td>lillie.test()</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1</td>
</tr>
<tr>
<td>PearsonTest()</td>
<td>pearson.test()</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1</td>
</tr>
<tr>
<td>ShapiroFranciaTest()</td>
<td>sf.test()</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1</td>
</tr>
<tr>
<td>CramerVonMisesTest()</td>
<td>ad.test()</td>
<td>0.9451477</td>
<td>0.9105691</td>
<td>1</td>
</tr>
<tr>
<td>CramerVonMisesTest()</td>
<td>lillie.test()</td>
<td>0.7299578</td>
<td>0.5339506</td>
<td>0</td>
</tr>
<tr>
<td>LillieTest()</td>
<td>cvm.test()</td>
<td>0.5339506</td>
<td>0.7299578</td>
<td>0</td>
</tr>
<tr>
<td>LillieTest()</td>
<td>ad.test()</td>
<td>0.5339506</td>
<td>0.7032520</td>
<td>0</td>
</tr>
<tr>
<td>Eps()</td>
<td>sf.test()</td>
<td>0.5333333</td>
<td>0.6315789</td>
<td>0</td>
</tr>
<tr>
<td>ShapiroFranciaTest()</td>
<td>ad.test()</td>
<td>0.7719298</td>
<td>0.3577236</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 9: Similarity report (the top 10 results) for the comparison between the code-base of the DescTools and nortest packages.

Discussion. We observe that 5 function pairs were marked as similar (Decision = 1). The top 4 results accurately indicate the corresponding normality tests from the two packages – their sources are identical.

However, the 5th result is a false positive: CramerVonMisesTest() is reported as similar to ad.test(), which implements the Anderson–Darling normality test. Let us “visually” compare their sources:

CramerVonMisesTest <- function(x) {
  DNAME <- deparse(substitute(x))
  x <- sort(x[complete.cases(x)])
  n <- length(x)
  if (n < 8)
stop("sample size must be greater than 7")
p <- pnorm((x - mean(x))/sd(x))
W~<- 1/(12 * n) + sum((p - (2 * seq(1:n) - 1)/(2 * n))^2))
WW <- (1 + 0.5/n) * W
if (WW < 0.0275) {
  pval <- 1 - exp(-13.953 + 775.5 * WW - 12542.61 * WW^2)
}
else if (WW < 0.051) {
  pval <- 1 - exp(-5.903 + 179.546 * WW - 1515.29 * WW^2)
}
else if (WW < 0.092) {
  pval <- exp(0.886 - 31.62 * WW + 10.897 * WW^2)
}
else if (WW < 1.1) {
  pval <- exp(1.111 - 34.242 * WW + 12.832 * WW^2)
}
else {
  warning("p-value is smaller than 7.37e-10,
  cannot be computed more accurately")
  pval <- 7.37e-10
}
RVAL <- list(statistic = c(W = W), p.value = pval,
  method = "Cramer-von Mises normality test",
  data.name = DNAME)
class(RVAL) <- "htest"
return(RVAL)
}
ad.test <- function(x) {
  DNAME <- deparse(substitute(x))
x <- sort(x[complete.cases(x)])
n <- length(x)
if (n < 8)
  stop("sample size must be greater than 7")
logp1 <- pnorm((x - mean(x))/sd(x), log.p = TRUE)
logp2 <- pnorm(- (x - mean(x))/sd(x), log.p = TRUE)
h <- (2 * seq(1:n) - 1) * (logp1 + rev(logp2))
A~<- -n - mean(h)
AA <- (1 + 0.75/n + 2.25/n^2) * A
if (AA < 0.2) {
  pval <- 1 - exp(-13.436 + 101.14 * AA - 223.73 * AA^2)
}
else if (AA < 0.34) {
  pval <- 1 - exp(-8.318 + 42.796 * AA - 59.938 * AA^2)
}
else if (AA < 0.6) {
  pval <- exp(0.9177 - 4.279 * AA - 1.38 * AA^2)
}
else if (AA < 10) {
  pval <- exp(1.2937 - 5.709 * AA + 0.0186 * AA^2)
}
else pval <- 3.7e-24
RVAL <- list(statistic = c(A = A), p.value = pval,
  method = "Anderson-Darling normality test",
  data.name = DNAME)
class(RVAL) <- "htest"
return(RVAL)
}

Basically, the main difference between the two functions is in the numeric constants used, which
we know is a kind of detail purposefully ignored by SimilaR. Indeed, knowing that the Anderson–Darling
test is a generalisation of the Cramér–von Mises test (both of them are based on the $L_2$-distance
between the empirical and true cumulative distribution function, the former is a weighted version
of the latter), makes the reported decision concerning the similarity judgement quite justified.

Interestingly, DescTools does provide the AndersonDarlingTest function, but this is a version of
the goodness-of-fit measure to test against any probability distribution provided. In other words, a c.d.f. of a normal distribution is not hard-coded in its source, and thus is significantly different from the code of ad.test().

It is also worth noting that there are no false positives in terms of statistical tool types – all the functions deal with some form of goodness-of-fit testing and we recall that DescTools defines ca. 550 functions in total.

**Discussion**

We have introduced an algorithm to quantify the similarity between a pair of R source code chunks. The method is based on carefully prepared Program Dependence Graphs, which assure that semantically equivalent code pieces are represented in the same manner even if they are written in much different ways. This makes the algorithm robust with respect to the most typical attacks. In a few illustrative examples, we have demonstrated typical code alterations that the algorithm is invariant to, for instance, aliasing of variables, changing the order of independent code lines, unwinding nested function calls, etc.

In the presented case study we have analysed the similarities between the DescTools and nortest packages. Recall that most of the cloned function pairs are correctly identified, proving the practical usefulness of the SimilaR package. The reported above-threshold similarity between CramerVonMisesTest() and ad.test() is – strictly speaking – a false positive, nevertheless our tool correctly indicates that the two functions have been implemented in much the same way. This might serve as a hint to package developers that the two tests could be refactored so as to rely on a single internal function – de-duplication is among the most popular ways to increase the broadly conceived quality of software code.

On the other hand, the algorithm failed to match the very much-different (implementation-wise) AndersonDarlingTest() (generic distribution) with its specific version of ad.test() (normal distribution family). However, comparisons of such a kind, in order to be successful, would perhaps require the use of an extensive knowledge-base and are of course beyond the scope of our tool.

Finally, let us note that due to the use of a polynomial-time algorithm, assessing the similarity of two Program Dependence Graphs is relatively fast. This makes SimilaR appropriate for mining software repositories even of quite considerable sizes. However, some pre-filtering of function pairs (e.g., based on cluster analysis) to avoid performing all the pairwise comparisons would make the system even more efficient and scalable.

Future versions of the SimilaR package will be equipped with standalone routines aiming at improving the quality and efficiency of R code, such as detecting dead or repeated code, measuring cyclomatic complexity, checking if the program is well structured, etc.

**Acknowledgements.** The authors would like to thank the Reviewers for valuable feedback that helped improve this manuscript.

**Bibliography**


B. Weisfeiler and A. A. Lehman. A reduction of a graph to a canonical form and an algebra arising during this reduction. Nauchno-Technicheskaya Informatsia, 2(9):12–16, 1968. [p7]


Maciej Bartoszuk  
Faculty of Mathematics and Information Science,  
Warsaw University of Technology  
uł. Koszykowa 75, 00-662 Warsaw, Poland  
https://bartoszuk.rexamine.com  
https://orcid.org/0000-0001-6088-8273  
m.bartoszuk@mini.pw.edu.pl

Marek Gagolewski  
School of Information Technology,  
Deakin University  
Geelong, VIC, Australia  
and  
Faculty of Mathematics and Information Science,  
Warsaw University of Technology  
uł. Koszykowa 75, 00-662 Warsaw, Poland  
https://www.gagolewski.com  
https://orcid.org/0000-0003-0637-6028  
m.gagolewski@deakin.edu.au