



北京大学

博士研究生学位论文

题目： 测试文档

姓 名： 某某

学 号： 0123456789

院 系： 某某学院

专 业： 某某专业

研究方向： 某某方向

导 师： 某某教授

某年某月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则一旦引起有碍作者著作权之问题，将可能承担法律责任。

摘要

pkuthss 文档模版最常见问题:

在最终打印和提交论文之前，请将 *pkuthss* 文档类选项中的 **colorlinks** 改为 **nocolorlinks**，因为图书馆要求电子版论文的目录必须为黑色，且某些教务要求打印版论文的文字部分为纯黑色而非灰度打印。

\cite、**\parencite** 和 **\supercite** 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记：**test-en**，**[test-zh]**、**[test-en, test-zh]**。

若要避免章末空白页，请在调用 *pkuthss* 文档类时加入 **openany** 选项。

如果编译时不出参考文献，请参考 **texdoc pkuthss** “问题及其解决”一章“其它可能存在的问题”一节中关于 **biber** 的说明。

关键词：其一，其二

Test Document

Test (Some Major)

Directed by Prof. Somebody

Abstract

Test of the English abstract.

Keywords: First, Second

目录

序言	1
0.1 Introduction	1
0.2 Problem	8
0.2.1 The C Preprocessor	8
0.2.2 Designing the Backward Transformation	10
0.2.3 Naive Solutions	12
0.3 Approach	14
0.3.1 Modeling forward preprocessing	14
0.3.2 Modeling the changes	17
0.3.3 Rewriting Steps	17
0.3.4 Backward Transformation with Replacement	18
0.3.5 Extending to full CPP	23
0.3.6 Extending to other types of changes	23
0.4 Evaluation	25
0.4.1 Research Questions	25
0.4.2 Setup	25
0.4.3 Threats to Validity	28
0.4.4 Results	28
0.5 Related Work	30
0.5.1 Bidirectional Transformation	30
0.5.2 Analyzing and editing unpreprocessed C code	30
0.5.3 Empirical studies on the C preprocessors	31

0.6 Conclusion	32
参考文献	33
附录 A 附件	39
致谢	41

序言

pkuthss 文档模版最常见问题:

在最终打印和提交论文之前, 请将 *pkuthss* 文档类选项中的 `colorlinks` 改为 `nocolorlinks`, 因为图书馆要求电子版论文的目录必须为黑色, 且某些教务要求打印版论文的文字部分为纯黑色而非灰度打印。

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记: `test-en`, `[test-zh]`、`[test-en, test-zh]`。

若要避免章末空白页, 请在调用 *pkuthss* 文档类时加入 `openany` 选项。

如果编译时不出参考文献, 请参考 `texdoc pkuthss` “问题及其解决”一章“其它可能存在的问题”一节中关于 `biber` 的说明。

0.1 Introduction

现在, 许多程序分析工具都涉及代码修改功能。在这些工具中, 有许多都是代码修复工具 [1–4]。通常来说, 修复工具的输入是一段代码和一组测试, 并不断修改代码直至代码能通过测试。另一些程序分析工具是API升级工具 [5–7]。当API升级时出现了不兼容情况时, 这些工具可以自动更新相应的API调用让程序与API契合。我们把这类直接修改代码的工具称作程序编辑工具。

A lot of program analysis tools involve direct modification of source code. A notable category of such tools is program-repair tools [1–4]. These tools take a program and a set of tests as input, and modify the program until all tests pass. Another category is API evolution tools [5–7]. When an API is upgraded with incompatible changes, these tools automatically change the API invocations to comply with the new API. We call such tools that directly modify the source code *program-editing tools*.

另一方面，许多程序语言的实现都带有预处理器 [8–10]。最常见的预处理器是C预处理器（C++）。许多程序语言也接受C++，包括C，C++和Objective-C。同时，程序员也时常使用C++来写一些零散的小工具。这时就会使用到预处理器。例如Korpela [11]曾在文章中描述过用C++写一个HTML编辑工具：这个工具会把页面间相同的HTML代码转换成C的宏，而不是直接生成HTML页面。然后页面再利用这些宏最终生成HTML文件。

On the other hand, many programming languages are provided with preprocessors [8–10]. The most widely used preprocessor is the C preprocessor (CPP). Many programming languages adopt CPP, notable C, C++, and Objective-C. Furthermore, CPP is often used by programmers in casual situations as a general purpose tool, where the preprocessor is added as a building step for any language used. For example, Korpela [11] describes the use CPP as an HTML authoring tool: instead of writing HTML pages directly, the shared code pieces between HTML pages are first defined as C macros, and HTML pages using these macros are preprocessed into final HTML files.

程序编辑工具通常不会去修改程序的预处理指令。但是，只有能够把修改映射到预处理之前的代码的工具才算有用。只在预处理后的代码中修复错误会导致原有程序再次编译的时候错误依然存在，这样毫无意义。这个问题具有挑战性，因为工具必须同时能理解预处理命令和目标程序语言，同时保证修改在两边能保持一致。事实上，现有的程序编辑工具往往无法正确处理预处理指令、或是直接不处理预处理指令，例如现有的C语言工具：GenProg [1, 2]，RSRepair [3]，和SemFix [4]。这三个工具都只在预处理后的代码上工作。用户需要手动检查预处理后的代码变化，并自行修改源代码——而这又增加了新bug的可能。

Program editing tools usually do not directly change preprocessor directives. However the tools must be able to map results back to the unprocessed source to be useful. There is no point of fixing a bug in the preprocessed code and only to have it overwritten when the unchanged source is compiled again. This is challenging, as the tools must be able to understand both the preprocessor directives and the target programming languages, and make sure whatever changes made on both levels are consistent with each other. As a matter of fact, existing program-editing tools often fail to produce correct results under the presence of preprocessor directives, or give up dealing with them entirely. We have in-

investigated the implementations of three influential bug-fixing tools on the C programming language: GenProg [1, 2], RSRepair [3], and SemFix [4]. All the three implementations work only on preprocessed code. Users have to manually inspect the preprocessed code, and copy the changes to the original source code—risking of introducing new bugs in the process.

代码重构是一个密切相关的领域 [12, 13]。在代码重构中，程序编辑工具有时需要直接修改预处理指令。比如：用户有时想重命名一个宏，或者需要提取一个宏作为重构的一部分。在这种情况下，工具开发者别无选择，只好修改预处理指令。典型情况中，工具开发者会定义一种新的C语法使得原有的C语法和预处理指令能兼容。但是，如果我们考虑更一般的程序编辑工具，这种方法就捉襟见肘了。首先，工具开发者需要在真正设计工具之前把精力花费在学习语言的细节上。其次，学习新语言的努力并不能在其他语言中复用。

A closely related area is refactoring [12, 13], where tools are expected to directly manipulate preprocessor directives. For example, one may well want to rename a macro or extract a macro as part of the refactoring. In this case, tool builders have no choice but to bite the bullet and confront the preprocessor directly. Typically a new C grammar is designed such that it incorporates both the original C grammar and the preprocessor directives. However, when applied to a more wider range of code editing tools, such almost brute force approaches exhibit obvious shortcomings. First, tool developers using such a grammar basically have to start from scratch: they have to learn the new grammar and leave behind the existing tool chains on C. Second, the effort spend on the new grammars is specialized and cannot be reused for other languages, which basically rules out casual uses of CPP.

本文中我们提出了一个轻量级的支持C++的程序编辑工具实现方法。该系统时一个双向的C++预处理器：原有的预处理过程可以背看作一个正向变换，在此基础上我们添加一个能够把预处理后代码上的修改反响映射回去的反向变换。于是，程序编辑工具现在可以关注于预处理后的代码，并把映射修改的交给我们的自动工具¹。

In this paper we propose a lightweight approach to support CPP in program-editing

¹ 这个过程并不是全自动的。因为我们的工具现在支持程序编辑的基本操作。尽管这些步骤可以用通用代码差分方法实现 [14]，但是如果工具能直接提供基本编辑操作可以有最好的效果

tools. Our system acts as a bidirectional CPP: the original preprocessing can be considered as a forward program transformation, and we add to it a corresponding backward transformation that maps changes on the preprocessed code back into the unpreprocessed source. As a result, program-editing tools can now focus on preprocessed programs, and have results (almost) automatically reflected to the source².

在这里我们列举一些例子：（1）上文中所提到的三个学术界认可的错误修复系统现在可以处理预处理前的代码；（2）API升级软件现在可以在有预处理代码的情况下更好地实现；（3）所有并不需要关心预处理过程的程序编辑工具都能够被改善。

We list a few examples here: (1) as mentioned above the implementations of the three state-of-the-art bug-fixing approaches only deal with preprocessed code; (2) the API evolution tools mentioned previously can also be implemented more conveniently by only dealing with preprocessed code; (3) all program-editing tools on languages that do not formally rely on CPP naturally fall into this category because the programs may be put under casual uses of CPP.

虽然现在存在着几种双向变换的技术 [15–17], 但是它们都是为数据的转换设计的。给定一个源数据集 s , 一个变换程序 p , 和一个目标数据集 $t = p(s)$, 这些方法试图将 t 上的变化描述成 s 的变化。然而, C的预处理器与数据的转换不同, 因为C的源代码不仅包含了作为数据的代码, 还包含了座位变换程序的预处理指令。这就要求反向变换的机制能处理更复杂的情况: 当目标数据发生变化时, 我们可能要变化源数据、转换程序、或者是二者都变化。

Although there exist several bidirectionalization techniques [15–17], they are designed for data transformation: given a source data set s , a transformation program p , and a target data set $t = p(s)$, these approaches try to reflect the changes on t to s . The C preprocessor differs from this data transformation scenario in the way that the unpreprocessed source program actually contains both the data set and the transformation program. This added complication amplifies the variance of the backward transformation: when the target data is changed, we may change either the source data set, the transformation

² It is not entirely automatic because the tools need to support the extraction of the changes on the preprocessed programs. Although this step can be performed by generic code differencing [14], extracting the changes directly from the tools gives the best result.

program, or both.

本系统的一个创新处是在能够处理上文提到的复杂双向变换情况的同时，尽量最小化对变换程序`p`的修改。首先，该设计中从不引入新的宏定义或修改现有的宏定义，有效地限制反向变化的影响。其次，该设计只会移除/修改宏调用而并不会创造新的宏调用。再有，该设计只会在必要的时候移除宏调用。这样，我们就能尽可能保证源代码和修改之前的相似度。

A key design novelty in our approach that serves to control this variance is to allow, but at the same time minimize, changes to the transformation programs. First, our approach never introduces new macro definitions or modifies existing macro definitions, effectively confining the impact of the reflected changes to a local scope. Second, our approach only removes existing macro invocations but never invent new ones. Furthermore, we will consider removing macro invocations only when necessary. In this way, we maintain the existing structure of the original program source as much as possible.

实现这样的算法设计也是十分具有挑战的。典型的这类双向变换的方法 [15, 16, 18] 是顺着程序的抽象语法树AST拆分双向变换。每一个子树对应着一个小的双向变换，组合起来就成为整个程序的双向变换。但是，一个未预处理的C++的程序并不能简单地解析成树状结构。比方说，在下列代码中，

```
#define inc(x) 1+x
#define double(x) 2*x
inc(double) 2
inc(double) (x)
```

第一个宏调用`inc(double)`会自动展开成为一个单独的**statement**。但是第二个宏调用不能单独展开成一个**statement**。它需要连上之后的`(x)`，循环展开后才算完整的句子。因此，我们不能把第二个宏调用`inc(double)`当作独立的一段并直接对他做双向变换的分析。为了克服这个困难，我们为描述类似C++程序这样的情况设计了新的模型。我们把预处理看作是对代码数据的重写规则（*rewriting rules*）集，而不是直接把代码解析成抽象语法树。这样的模型把程序的双向变换看作是重写规则的双向变换。

Implementing this design is also challenging. Typical approaches to bidirectionalization [15, 16, 18] decompose the transformation along the abstract syntax tree of the program, where each subtree corresponds to a small bidirectional transformation that

collectively forms the final transformation. However, a CPP program cannot be easily parsed into a tree structure. For example, in the following piece of code,

```
#define inc(x) 1+x
#define double(x) 2*x
inc(double) 2
inc(double) (x)
```

the first `inc(double)` independently expands to a new segment, but the second `inc(double)` is only part of an expansion, as the expanded `double` will form a new macro invocation with `(x)` to be recursively invoked. As a result, we cannot treat `inc(double)` as an independent unit and derive a backward transformation from it. To overcome this difficulty, we propose a new model for interpreting CPP programs. Instead of parsing a CPP program into an abstract syntax tree, we view the preprocessing as applying a set of rewriting rules to the code. This interpretation enables the bidirectionalization of a CPP program as bidirectionalization of each rewriting rule.

另外，该设计也可以被证明满足双向变换的正确性：（1）如果预处理后的程序没有变化，那么源代码也不会变化；（2）源代码在接受了反向映射的变化后，预处理后会得到相同的变化后的预处理后的代码。这两个性质在双向变化领域被称作GETPUT和PUTGET [19]。

Furthermore, our approach is proved to be correct, in the sense of the following round-trip laws (1) if the preprocessed program is not changed, the unpreprocessed program will not be changed, and (2) preprocessing the unpreprocessed program with the reflected changes will produce exactly the same changed preprocessed program. These two properties are known as GETPUT and PUTGET [19] in the bidirectional transformation literature.

总之，该项目的贡献如下：

- 我们提出了一个轻量级的在程序编辑工具方面的双向C预处理器。我们分析了不同可能的设计并提出了五个反向变换应有的性质，包括GETPUT和PUTGET (Section 0.2)。
- 我们提出了一个能够符合五个性质的算法。该算法把C的预处理看作是重写规则的集合，并结构性地把预处理的反向变换转换到重写规则的双向变换上 (Section 0.3)。

- 我们在Linux内核上验证了我们的方法，并和另外两种基本的做法进行比较：一个是直接把整个修改的文件映射回去；另一个是只把修改了的代码行映射回源代码。实验的结果显示相较于其他方法，我们的方法破坏了相当少的宏调用，并且总是可以给出正确的修改，而其他方法有时不行(Section 0.4)。

To sum up, our paper makes the following contributions:

- We propose a lightweight approach to handling the C preprocessor in program-editing tools based bidirectional transformations. We analyze different design alternatives and propose five requirements for defining the behavior of the backward transformation, including GETPUT and PUTGET (Section 0.2).
- We propose an algorithm that meets the five requirements. This algorithm is based on an interpretation of CPP as a set of rewriting rules, which structurally decomposes the bidirectionalization of CPP into the bidirectionalization of each rule
- We evaluate our approach on the Linux kernel and compare our approach with two baseline approaches: one reflecting changes by copying back the entire changed file and one reflecting changes by copying back the changed lines. The evaluation results show that our approach breaks much less macro invocations, and always produces correct results while the other two do not

Finally, we discuss related work in the paper in

0.2 Problem

0.2.1 The C Preprocessor

表 1: Main preprocessor directives and operators

Directives	Functionality	Example	Result
<code>#pragma</code>	Compiler options	<code>#pragma once</code>	removed from the preprocessed file
<code>#include</code>	File Inclusion	<code>#include <stdio.h></code>	the content of "stdio.h"
<code>#if</code> , <code>#ifdef</code> , ...	Conditional compilation	<code>#ifdef FEATURE1</code> <code> x = x + 1;</code> <code>#endif</code>	<code>x = x + 1;</code>
<code>#define X</code>	Object-like macro definition	<code>#define X 100</code> <code>a = X;</code>	<code>a = 100;</code>
<code>#define X(a, b)</code>	Function-like macro definition	<code>#define F(x) x * 100</code> <code>F(10);</code>	<code>10 * 100;</code>
<code>a ## b</code>	Concatenation	<code>#define X a_##100</code> <code>X</code>	<code>a_100</code>
<code>#b</code>	Stringification	<code>#define F(x) #x;</code> <code>F(hello);</code>	<code>"hello";</code>
<code>__FILE__</code> , <code>__DATE__</code> , ...	Predefined macros	<code>__FILE__</code>	<code>main.c</code>

Table 1 shows the main preprocessor directives and operators. A preprocessor directive starts with a `#` at the beginning of the line and ends at the end of the line. Macros can be defined by preprocessor directives but themselves are not preprocessor directives. Basically, we have four main types of preprocessor directives: `#pragma` providing compilation options, `#include` for including header files, `#if` for conditional compilation, and `#define` for macro definitions. Additionally, within a macro definition, we can use operators such as `##` and `#`, for concatenating two variables or quoting a variable. Finally, there are some pre-defined macros such as `__FILE__`, which will be replaced by the current values of the fields.

When the C preprocessor processes a source file, it transforms the source file according to the following rules:

- The `#include` and `#if` directives are first expanded, and then the expanded token sequences are scanned.
- For each macro invocation, the arguments are first preprocessed, and then the

invocation is expanded.

- If an argument contains # or ##, the unpreprocessed argument is used, otherwise the preprocessed argument is used.
- After a macro invocation is expanded, the expanded token stream is scanned again, where any newly introduced macro invocations are again expanded.
- To avoid infinite recursive expansion, if a macro has been expanded during the expansion process, it will not be expanded again.
- Any new preprocessor directives produced in the expansions will not be used in processing.

```
#define SAFE_FREE(x) if (x) vfree(x);
#define FREE(x) vfree(x);
#define RESIZE(array, new_size, postprocess) \
    g_resize_times++; \
    postprocess(array); \
    array = vmalloc(sizeof(int)*(new_size));
#define GARRAY(x) g_array##x;

RESIZE(GARRAY(2), 100, FREE);
```

图 1: An example of code preprocessing

As a concrete example, let us consider the code in Figure 1. This example is contrived to exhibit many real world patterns of macro uses. This piece of code contains four macro definitions and a macro invocation. The first two macro definitions are wrappers for a user-defined free function. The third one is for resizing an array with user-defined memory management and logging. The last one is for naming a set of special global variables. When the code is scanned by the preprocessor, first the arguments of RESIZE are processed. In this case, GARRAY(2) is expanded into g_array2. Though the third parameter FREE has been defined as a function-like macro, no argument is provided for FREE, so the preprocessor does not treat it as a macro invocation at this stage. Then the macro invocation to RESIZE is expanded, leading to the following code:

```
g_resize_times++;
FREE(g_array2);
g_array2 = malloc(sizeof(int)*(100));
```

Now we can see that in the expanded macro, an argument list is actually provided for `FREE`. Then the system will expand `FREE(g_array2)`, resulting in the following code.

```
g_resize_times++;  
vfree(g_array2);  
g_array2 = malloc(sizeof(int)*(100));
```

In other words, `RESIZE` is in fact a high-order macro where its third parameter is also a macro.

0.2.2 Designing the Backward Transformation

Now let us consider that a program-editing tool takes the preprocessed code. Realizing the invocation to `vfree` is unsafe, the programming editing tool inserts a guard before it, as shown below.

```
g_resize_times++;  
if (g_array2) vfree(g_array2);  
g_array2 = malloc(sizeof(int)*(100));
```

Now let us consider how to design a backward transformation that takes this change as input, and produces a change on the unpreprocessed code, where the changed unpreprocessed code will produce exactly the changed preprocessed code. There are two options: 1. we can change the definition of `RESIZE` macro, adding this guard into the macro definition. 2. we can expand the invocation to `RESIZE` macro, and add the guard to the expanded form. The first option makes global changes, affecting all places that invoke `RESIZE`. The second option makes only local changes, affecting only one place. However, since the backward transformation do not have the knowledge whether the change should be global or local, it is safer to choose the local option. In this example, probably we do not need to add the guard to every invocation to `vfree`, as it will cause unnecessary runtime overhead. The program-editing tool will have the knowledge to determine the safety of `vfree` calls and add this guard to all necessary places. Furthermore, the local option also minimizes the interruption to the original program, as the global option affect many places in the program. This consideration forms our first requirement of the backward transformation.

Requirement 1 *A backward transformation should not change any macro definition.*

According to the requirement, we should expand the macro invocation with the guard added. Then there are again two options. We may subsequently expand all macro expansions as follows.

```
g_resize_times++;
if (g_array2) vfree(g_array2);
g_array2 = malloc(sizeof(int)*(100));
```

(1)

Or we may expand only one level as follows.

```
g_resize_times++;
if (GARRAY(2)) FREE(GARRAY(2));
GARRAY(2) = malloc(sizeof(int)*(100));
```

(2)

We think code piece (2) is better than code piece (1) because it keeps more of the original code structure that are easier to understand, reuse, and maintain. This leads to our second requirement.

Requirement 2 *A backward transformation should aim to keep existing macro invocations.*

Some might be tempted to go a step further by abstracting the guard into a new macro, or trying to reuse an existing macro that covers its functionality, as in the following.

```
g_resize_times++;
SAFE_FREE(GARRAY(2));
GARRAY(2) = malloc(sizeof(int)*(100));
```

However, this is dangerous as very often macro definitions are not designed to replace all occurrences of its expanded form. For example, Ernst et al. 8 reports a macro definition, `#define ISFUNC 0`, that defines a constant used in certain system calls. It obviously does not mean we shall replace all occurrences of `0` with `ISFUNC`. This leads to our third requirement.

Requirement 3 *A backward transformation should not introduce new macro invocations.*

Besides the three requirements, we have two additional correctness laws borrowed from the bidirectional transformation literature, namely GETPUT and PUTGET 19. Let

s be the unpreprocessed program, t be the preprocessed program, c_t be the change to t , and c_s be the change to s that is produced by a successful backward transformation.

Requirement 4 (GETPUT) *If c_t is empty, then c_s is empty.*

Requirement 5 (PUTGET) *Let $s' = c_s(s)$ be the new unpreprocessed program obtained by applying c_s to s , and $t' = c_t(t)$ be the changed preprocessed program. Preprocessing s' produces t' .*

The requirements together define the behavior of a backward transformation: it should propagate the change back into the unpreprocessed code by means of modifying macro parameters, modifying normal text, and expanding macro invocations, and at the same time expanding as fewer macro invocations as possible.

0.2.3 Naive Solutions

Naive solutions does not meet our above listed requirements as we demonstrate in this subsection.

Naive solution I (per-file). The first naive solution is to directly copy back the changed preprocessed files and replace the original unpreprocessed files. This solution is the easiest to implement but has two major deficiencies. First, the original unpreprocessed source files may contain macro definitions, which will be lost if we directly copy back the preprocessed files, causing problems elsewhere. Second, this will leads to the expansion of all macro invocations in the source file, as well as all `#include` directives, destroying modularity completely.

Naive solution II (per-line). The second solution is to use more fine-grained units, copying back only the changed lines. This is doable with the assumption that modern compilers keep a traceability relationship between the lines in the unpreprocessed files and the preprocessed files. For example, when GCC preprocess our running example, it will replace lines 1-7 into empty lines, and put the three statements expanded from line 9 in one line, so that a one-to-one correspondence between lines are kept.

Although the second solution has the benefit of not removing macro definitions, it still has problems. First, a lot of macros are still unnecessarily expanded. In our example, if we copy back the changed line, code piece (1) will be the result, which goes against our second requirement and destroys modularity. The situation becomes even worse when we consider tools that copy lines of code within source files, e.g., GenProg 1 copies statements between two positions in the source files to fix bugs. In such cases all macro invocations in the copied lines will be lost. Second, if the original macro invocations span several lines, the backward mapping produces wrong results. For example, suppose there is a line break within the original macro invocation, as follows.

```
RESIZE(GARRAY(2),  
      100, FREE);
```

In GCC, the macro invocation will be expanded into two lines, where the first line contains the three expanded statements, and the second line is empty. As a result, in the backward transformation only the first line will be copied back, resulting in an incorrect program.

0.3 Approach

As mentioned before, the basic idea of our approach is to interpret CPP as a set of rewriting rules, where the backward transformation reverses each application of the rules. In this section, we first present this interpretation of CPP (Section 0.3.1). Then we describe the change operations we support and first focus on one type of change: replacements. (Section 0.3.2). Since we need to reverse each application of the rules, we first record the applications as rewriting steps (Section 0.3.2), and then derive a backward transformation for each rule, respectively (Section 0.3.3). We also discuss how this process satisfies the requirements (Section 0.3.4) and an optimization of the algorithm (Section 0.3.4). Finally, we discuss how to convert other types of changes into replacement (Section 0.3.6).

For the ease of presentation, we shall consider a subset of C preprocessor first: there is no `#` operator nor `##` operator, no `#include` directive, nor does any macro expansion contain any self invocation. We will later discuss how to extend this simplified model to a full model (Section 0.3.5).

0.3.1 Modeling forward preprocessing

We view a CPP program as a sequence of tokens. To recognize preprocessor directives from the sequence of tokens, we rely on two special tokens: `#` at the beginning of a line, and the end of line following from a `#` at the beginning of the line. We also assume all macro definitions in scope are stored in a *context*.

We deem the semantics of CPP as a set of rewriting rules. Each rewriting rule is taken the form of `guard` \hookrightarrow `action`. When `guard` is true, `action` is performed to replace some tokens at the beginning of the current remaining token sequence, and set the position for the next rule application.

Both `guard` and `action` are modelled as functions. Function `guard` takes the currently remaining token sequence and the current context, and returns a Boolean value to indicate whether the rule can be applied to the current token sequence or not. Function `action` takes the currently remaining sequence and the current context, and returns four components (`finalized`, `changed`, `restIndex`, `newContext`). This result indi-

cates that the tokens before `restIndex` are replaced by two consecutive subsequences, `finalized` and `changed`, where `finalized` is a sequence that does not need to be further scanned, and `changed` is a sequence that needs to be scanned. The last component, `newContext`, is the updated context. The algorithm for forward preprocessing is given in Algorithm 1. It repeatedly applies the rules in `R` until the whole token sequence has been processed.

```

Input: token sequence src, rule list R
Output: new token sequence res
ctx  $\leftarrow$  {};
while src.length  $>$  0 do
    for r  $\in$  R do
        if r.guard(src, ctx) then
            (finalized, hanged, rest, ctx')  $\leftarrow$  r.action(src, env);
            break;
        end
    end
    res  $\leftarrow$  res + finalized;
    src  $\leftarrow$  changed + src.sub(rest);
    // sub(1) returns a subsequence starting from 1
    ctx  $\leftarrow$  ctx';
end

```

Algorithm 1: Algorithm for forward preprocessing

There are totally four rules in the rule list `R`. One processes conditionals such as `#if`, `#ifdef`, one processes all other preprocessor directives as we basically need to remove them and updates the context, one processes macro invocations, and the last one processes normal text. These rules are described below.

- **R1:** This rule processes conditional compilations. Function `guard` determines whether the token sequence starts with a respective directive such as `#if`, `#ifdef`. Function `action` first evaluates the condition based on the current context, and then replaces the whole conditional with either the true branch or the false branch. The whole replacing sequence is returned as `changed` and `finalized` is empty.
- **R2:** This rule processes other preprocessor directives. Function `guard` determines whether the token sequence starts with a preprocess directive. Function `action` parses the directives, makes necessary changes to the context, and returns the index

of the first token after the directive as `restIndex`, with `finalized` and `changed` empty.

- R3: This rule expands a macro invocation. Function `guard` determines whether the first token in the sequence is the name of an object-like macro or the first two tokens are the name of a function-like macro and an open parenthesis. Function `action` consists of two steps.
 - First, the arguments are preprocessed by recursively applying R3 and R4³. When expanding an argument, the current context is used, and the token sequence includes only the argument.
 - Second, the occurrences of the parameters in the macro body are substituted by the preprocessed arguments. The new body are returned as `changed`, with `finalized` empty, and `restIndex` is the index of the next token after the macro invocation.
- R4: This rule processes normal tokens not captured by the other rules. Function `guard` always returns true. Function `action` returns the first token as `finalized`, an empty `changed`, and the index of the next tokens as `restIndex`.

As an example, let us consider the following program.

```
#define x 100
hello x
```

(3)

The first applicable rule is R2, which parse the macro definition, store it in the context and moves to the next line. The remaining sequence is now `hello x`. Then R4 is applied to move `hello` into the `finalized` sequence. Next, R3 is applied to expand `x` into `100`. Finally, `100` is scanned again and R4 moves it into the `finalized` sequence. Now the remaining sequence is empty and the preprocess stops.

³ Including preprocessor directives in the argument list is an undefined behavior in C++ 20. Here we ignore directives in arguments

0.3.2 Modeling the changes

Program-editing tools can change the program with a variety of operations. In this paper we consider three basic kinds: replacement, insertion, and copy. These operations are summarized from popular bug-fixing tools such as GenProg 1 and SemFix 4, which typically fix a bug by copying or creating an expression to replace another one or inserting at some location.

All the three operations directly manipulate the token sequences. A replacement is a pair (l, s) , where l is the index of the token to be replaced, and s is a token sequence that will replace the token at l . An insertion is a pair (l, s) , where token sequence s will be inserted after the token at index l . A copy is a triple (l, l_b, l_e) , showing the token sequence between index l_b (inclusive) and index l_e (exclusive) is copied after the token at l .

Note that the replacement naturally subsumes deletion. A change $(l, [])$ deletes a token, where $[]$ denotes an empty sequence.

In the rest of the section we shall first show how to perform a backward transformation with replacements and then map the other two types of operations to replacements.

0.3.3 Rewriting Steps

As mentioned in the introduction, our backward transformation reverses each application of rewriting rules. To perform a backward transformation, we use a data structure to record what rules have been applied to what parts of the source. We call such a record *rewriting steps*. Each rewriting step is a quadruple, (src, i, ctx, r) , specifying that rule r has been applied to a sub sequence of src starting at i with context ctx . Token sequence src is always the current token sequence, containing all modifications made from previous rule applications. Accordingly, a forward transformation is recorded as a sequence of rewriting steps. For example, preprocessing code piece 3 produces the following sequence of rewriting steps: $(P, 0, \{\}, R2)$, $(\text{hello } x, 0, \{x\}, R4)$, $(\text{hello } x, 1, \{x\}, R3)$, and $(\text{hello } 100, 1, \{x\}, R4)$, where P is the original program.

0.3.4 Backward Transformation with Replacement

The basic idea of our backward transformation is to propagate the changes backward along the sequence of rewriting steps, such that if we perform a forward preprocessing again, either exactly the same rewriting step is performed, or this rewriting step is replaced by a new sequence of equivalent rewriting steps (whose length can be zero). The algorithm for backward transformation can be found in Algorithm 2. The function `backward` propagates changes along one rewriting step, or reports a failure if no proper propagation can be found.

```
Input: a sequence of rewriting step rs  
Input: a set of replacements c  
Reverse sequence r;  
for r  $\in$  rs do  
|   c  $\leftarrow$  backward(rs, c);  
|   if backward failed then  
|   |   print “Changes cannot be applied.”;  
|   |   return;  
|   end  
end  
return c;
```

Algorithm 2: Algorithm for backward transformation

The key for implementing the backward transformation is to implement the function `backward`. The behavior of `backward` differs for each rewriting rule, and we discuss it from the simplest to the most complex rule.

First, if the rewriting step is performed by R2, the forward preprocessing consumes the preprocessor directive. The backward transformation shifts the replacements according to the length of the consumed macro. For example, consider the following CPP program.

```
#undef hello  
hello
```

Now if we have a replacement $(0, x)$, i.e, `hello` is changed into `x`, the backward transformation shifts the replacement into $(4, x)$ to ensure that it still changes the macro invocation `hello`.

Second, if the rewriting step is performed by R1, the forward transformation replaces

the whole conditional directive with one of its branches. In the backward transformation, we map the changes on the replaced branch back to its original location, and shift other changes if necessary. Since we never change macro definitions, the conditional is guaranteed to evaluate to the same branch.

Third, if the rewriting step is performed by R4, the forward transformation moves one token to `finalized`. The backward transformation needs to ensure that R4 is still performed on the changed token sequence. For example, if the user replaces `hello` in `hello c` with `a b`, we need to check, starting from `a`, whether the only viable path is still to apply R4 twice to preprocess the replacing sequence, i.e., only guard of R4 evaluates to true on either `a b c` or `b c`. Here failures may occur. For example, if the user changes `hello` into `x`, where `hello` is not a macro but `x` is an object-like macro. Another example is that the user changes `GARRAY hello` into `GARRAY (hello)`, where `GARRAY` is a function-like macro. In this case, `GARRAY` was previously preprocessed by R4 because it is not followed by an open parenthesis, whereas in the changed sequence `GARRAY` is, and therefore the backward transformation fails at the rewriting step on `GARRAY`. Such failures are necessary as in both the above cases, there is no way to propagate back the changes.

Finally, the case of R3, which expands a macro invocation, is the most complex. R3 consists of two sub steps. We first try to reverse the two sub steps in an reversed order. If any of the sub step fails, we try to expand the macro invocation.

In the second sub step, the occurrences of parameters in a macro body are replaced by expanded arguments. The backward transformation maps the changes back to the arguments. There are two cases where an expansion may be triggered: (1) tokens that are not from the arguments are changed, and (2) multiple occurrences of the same parameter are changed to different values. For example, giving the following piece of code,

```
#define x 1
#define plus(a) a+a
plus(x);
```

(4)

if we changes `1 + 1` into `1 - 1` or `1 + 2`, an expansion will be triggered.

In the first sub step, the arguments are preprocessed. The backward transformation

recursively call `backward` to propagate the changes along the rewriting steps of each argument. Finally, a safety check is performed on the propagated changes: if the changed argument contains a comma at the top level (and the comma is not enclosed by a pair of parentheses) or any unmatched parenthesis, we trigger an expansion. This is because a comma or an unmatched parenthesis can break the original structure of arguments.

If the backward transformation on any of the sub step fails, we expand the macro invocation. The expansion will generate changes that replace the macro invocation with its expanded body. As an example, let us assume the second 1 in `1 + 1` is changed into 2 in code piece 4. After backward transformation, the propagated change will replace `plus` in `plus(x)` with `x + 2` and delete `(x)`.

The key to the expansion is how to construct the replacing sequence, `x + 2`. From the forward transformation we know that the unchanged preprocessed token sequence is `1 + 1`, and both 1s are from the argument `x`. So we copy the rewriting steps of `x` to the locations of the two 1s, and use the copied rewriting steps to propagate back the changes on the 1s. The first 1 is not changed, so the original `x` is kept. The second 1 is changed into 2, so the macro invocation `x` is also expanded, and we get the final text `x + 2`.

However, we cannot always copy the rewriting steps of an argument to the occurrences of its corresponding parameter. Basically, when we copy the rewriting step of an argument, we are assuming that the occurrence of the parameter can be replaced by its unprocessed argument, and all rewriting steps behave exactly the same as before. For example, when expanding `plus(x)`, we can replace the two occurrences of `a` by `x`, forming `x + x` where the two `x` expands exactly the same way as the argument `x`, forming `1 + 1`. However, this is not always the case. Let us consider the following code.

```
#define p (x)
#define pplus(x) plus x hello
pplus(p)
(5)
```

With the macros defined in code piece 4, the above code is preprocessed into `1+1 hello`. However, if we need to expand `pplus(p)`, we cannot expand it into `plus p hello`, because it will only expand to `plus (1) hello`. This is because the use of `p` instead `(x)` breaks

the original macro invocation.

As a result, we have to add a safety check when we copy rewriting steps. We copy the rewriting steps of an argument only if none of the following conditions is satisfied.

- The preprocessed argument starts with a left parenthesis.
- The preprocessed argument contains a top-level comma.
- The preprocessed argument contains unclosed parentheses.
- The preprocessed argument becomes different if we preprocess it again.

The first three conditions correspond to situations similar to the above example: the preprocessed argument is used as part of another macro invocation in the expanded form, and replacing it with the unpreprocessed one will break the macro invocation. The last condition corresponds to the following case.

```
#define id(x) x  
id(plus p)
```

This piece of code expands to $1 + 1$, but if we expand `id(plus p)` into `plus p`, the other macro expansions will be blocked. In other words, with macro invocation, an argument will be scanned twice, but when the macro invocation is expanded, an argument will only be scanned once. We need to make sure this does not affect correctness.

Correctness

We can directly reason that the backward transformation satisfies Requirements 1, 3, 4 and 5. Requirement 2 will later be evaluated by our experiments. Requirement 1 holds because we never propagate changes to macro definitions. Requirement 3 holds because if we happen to introduce a new macro invocation, we must have changed a normal token that should be preprocessed by R4 into part of a macro invocation, and the check in R4 will prevent such a change. Requirement 4 holds because we introduce new changes only when we expand a macro invocation, and we would not expand a macro invocation if nothing is changed. Requirement 5 has been reasoned throughout this subsection. To prove it formally, we need to inductively prove that each rewriting step either still behaves the same on the changed steps, or is replaced by a sequence of equivalent rewriting steps (the size of the sequence may be zero). A tricky point is when we copy rewriting steps of

arguments to the expanded method body, the copied rewriting steps may be mixed with later rewriting steps. We can show that, still by induction, the order of applying these rewriting steps can be changed without affecting the final result. We omit the details due to space limitations.

Optimization

In the current algorithm, we need to record the whole program after each rewriting step, even if only a small portion has changed. In the backward direction, we also need to shift all changes at each step. This contributes to a major performance penalty of the algorithm.

To optimize the algorithm, we first divide the original token sequence into a set of subsequences, where each subsequence is independently preprocessed by a set of rewriting steps. In this way, we can treat each subsequence as an independent program to perform the backward transformation, and then merge the changes. Given a rewriting step (src, i, ctx, r) , if the token at i is not generated by a preceding rewriting rule, i.e., is from the original program, then we say the location right before i is a *splitting point*. With this definition, we guarantee that no rule application crosses a splitting point. We divide the original program into subsequences along the splitting points, and perform the backward transformation independently for each subsequence before merging the changes. When merging changes from two subsequences, we need to check that the preprocessed version of left sequence and the unpreprocessed version of the right sequence would not form a new macro invocation. This check was performed by R4 in the unoptimized algorithm.

For example, code piece 3 can be divided into three subsequences: the macro definition, `hello`, and `x`. Suppose `plus` is a macro defined earlier. If the user changes `hello` into `plus` and changes `100` into `(x)`, a new macro invocation is formed across the boundary and we should report a failure.

0.3.5 Extending to full CPP

In this sub section we discuss how we can extend the above algorithm to support full CPP. Due to space limitations, we only discuss the main ideas without the full details.

To support `#` and `##` operators in forward preprocessing, we need to add two additional types of sub steps in R3, one for stringifying tokens and one for concatenating tokens. Furthermore, parameters used with these operators are replaced by their unpreprocessed forms but not their preprocessed forms, so in the first sub step we need to keep both the unpreprocessed and the preprocessed forms. We need to add three extensions to the backward transformation. First, we need to design the backward transformation for the two new sub steps. Second, we need to add a safety check to determine whether the unpreprocessed form and the preprocessed form are changed consistently. Finally, when expanding a macro, we should not try to recover `#` and `##` operators as the two operators cannot exist on the top level.

To support `#include` in the forward preprocessing, we need to add another rewriting rule to support `#include`. In the backward transformation, we need to trace how changes are propagated to each file, and check changes propagated back from different `#include` directives of the same file are consistent.

Finally, CPP does not allow expanding a macro within its own expansion to prevent infinite loops. To support this, we need to add finer control of the context in our forward rewriting rules. When we are dealing with the tokens expanded from an invocation to macro `m`, the definition of `m` should be removed from the current context.

0.3.6 Extending to other types of changes

We have discussed how to deal with replacements. Now let us proceed to insertions and copying. Note that an insertion can be directly converted into a replacement. If we insert a token `y` before a token `x`, we can convert it as replacing `x` with `y x`. However, this direct conversion may cause unnecessary macro expansions. For example, if we insert `y` before `100` in the preprocessed code piece 3, `hello 100`, we should not expand `x` in the backward transformation, but if we model the insertion as replacing `100` with `y 100`, our

backward transformation will expand x because its body has been changed.

The above example exhibit an insertion at a splitting point, and such an insertion is guaranteed not to expand macros. To reduce the number of unnecessary expansions, We treat the inserted token sequence at a splitting point as an independent sequence and check whether only R4 is applicable to it, since there is no way to put it back if any other rewriting rules can be applied on the inserted sequence. Then we use the same method used in the optimized algorithm to merge the subsequences.

The copy operation is similar to insertion. The only difference is that copied segments may contain macro invocations and we shall try to recover these macro invocations. For this, we perform a special backward transformation. First, we generate changes on the preprocessed code that deletes all tokens except the segments being copied. Then we perform a backward transformation by ignoring rewriting steps with R1, R2, and R3 whose associated macro is not defined or defined differently at the target position. In this way we ensure that only macros that is defined at the target positions are recovered. Then, we insert the token sequence returned by the backward transformation to the target position.

0.4 Evaluation

0.4.1 Research Questions

In this section we focus on the following research questions.

- **RQ1: Macro Preservation.** According to requirement 2, our approach aims to preserve existing macro invocations. How does the strategy perform on actual programs? How does it compare to other techniques?
- **RQ2: Correctness.** Our approach is guaranteed to be correct according to requirements 4 and 5. How important is this correctness? How does our approach compare to other techniques that do not ensure correctness?
- **RQ3: Failures.** Our approach may report a failure when it cannot find a proper way to propagate the change. How often does this happen? Are the failures false alarms (there exists a suitable change but our approach cannot find it)?

To answer these questions, we conducted a controlled experiment to compare our approach with the two naive approaches described in Section 0.2.3 on a set of generated changes on Linux kernel source code. In the rest of the section we describe the details of the experiment.

0.4.2 Setup

Implementation

We have implemented our approach in Java by modifying JCPP⁴, an open source C Preprocessor. We also implemented the two naive approaches in Section 0.2.3 for comparison. Our implementation and experimental data can be found on our web site⁵.

Benchmark

Our experiment was conducted on the Linux kernel version 3.19. We chose Linux source code because Linux kernel is one of the most widely used software projects

⁴ <http://www.anarres.org/projects/jcpp/>

⁵ <https://github.com/harouwu/BXCPP>

implemented in C. It contains contributions from many developers, and has a lot of preprocessor directives and macro invocations.

To conduct our experiment, we need a set of changes on the Linux kernel code. Since we concern about how different backward transformations affect preprocessing, we generated changes only in functions that contain macro invocations. To do this, we first randomly selected 180 macros definitions from the kernel code. Since there are far more object-like macros than function-like macros, we would select very few function-like macros if we use pure random selection. So we controlled the ratio between object-like and function-like macros to be 1.5 : 1. Based on the selected macros, we randomly selected a set of functions which contain invocations to the macros. Finally, we randomly selected 8000 lines from the functions. There are in total 133 macro invocations in the selected lines.

Next we generated a set of changes on the selected lines. To simulate real world changes, we randomly generated two types of changes. The first type is token-level change, in which we randomly replace/delete/insert a token. The second type is statement-level change, in which we delete a statement or copy another statement to the current location. These two types of changes are summarized from popular bug-fixing approaches [1](#), [3](#), [21](#). The statement-level changes are directly used by GenProg [1](#) and RSRepair³. The token-level changes simulate small changes such as replacing the argument of a method or change an operators used in approaches such as PAR [21](#).

More concretely, we had a probability p to perform an operation on each token, where the operation is one of insertion, replacement and deletion, which had equal probability. The replacement was performed by randomly mutating some characters in the token. The insertion was performed by randomly copying a token from somewhere else. Similarly, we had a probability q to perform an operation on each statement, where the operation is copy or deletion. The copied statement was directly obtained from the previous statement. We recognized a statement by semicolon.

Different tools may have different editing patterns: a migration tool typically changes many places in a program, whereas a bug-fixing tool may change a few places to fix a bug. To simulate these two different densities of changes, we used two different set of

probabilities. For the high-density changes, we set $p = 0.33$ and $q = 0.1$. For the low-density changes, we set $p = 0.1$ and $q = 0.05$.

We generated ten sets of changes, five with high-density and five with low-density. The number of the changes generated for each set is shown in Table 2.

表 2: Changes generated for the experiment

Low	Set	1	2	3	4	5
Density	Changes	952	885	956	967	884
High	Set	6	7	8	9	10
Density	Changes	3133	3136	3088	3123	3048

Independent variables

We considered the following independent variables. (1) *Techniques*, we compared our approach with the two naive solutions, per-file and per-line. (2) *Density of changes*, we evaluated both on the five high-density change sets and the five low-density change sets.

Dependent variables

We considered two dependent variables. (1) *Number of remaining macro invocations*. We re-ran the preprocessor after the backward transformation, and counted how macro invocations are expanded during preprocessing. Since none of the techniques will actively introduce new macro invocations, the number of expanded invocations is the number of remaining invocations. To avoid noise from included files, we count only the macro invocations in the current file. (2) *Number of errors*. We re-ran the preprocessor, and compared the new preprocessed program with the previously changed program by Unix file-comparing tool `fc`. Every time `fc` reported a difference, we counted it as an error. (3) *Failures*. Our approach may fail to propagate the changes, and we record whether a failure is reported for each change set.

0.4.3 Threats to Validity

A threat to external validity is whether the results on generated changes can be generalized to real world changes. To alleviate this threat, we used different types of changes and different density of changes, in the hope of covering a good variety of real-world changes.

A threat to internal validity is that our implementation of the three approaches may be wrong. To alleviate this threat, we investigated all errors we found in the experiments, to make sure it is a true defect of the respective approach but not a defect in our implementation.

0.4.4 Results

表 3: *Experimental Results*

Low Density	Set	1	2	3	4	5
Our Approach	Macros	73	75	72	80	81
	Errors	0	0	0	0	0
	Failures	n	n	n	n	n
Per-Line	Macros	23	25	23	20	26
	Errors	6	7	6	7	7
Per-File	Macros	0	0	0	0	0
	Errors	0	0	0	0	0
High Density	Set	6	7	8	9	10
Our Approach	Macros	47	51	53	48	44
	Errors	0	0	0	0	0
	Failures	n	n	n	n	n
Per-Line	Macros	9	7	7	8	10
	Errors	6	6	7	6	6
Per-File	Macros	0	0	0	0	0
	Errors	0	0	0	0	0

Row “Macros” shows the number of remaining macros. Row “Errors” shows the number of errors caused. Row “Failures” indicates whether a failure is reported in the backward transformation.

The result of our evaluation is shown in Table 3. We discuss the results with respect to the research questions below.

RQ1

As we can see, our approach preserves macro invocations. Per-line preserves very few macro invocations, while per-file, as we expected, preserves no macro invocations. We further investigated why per-line preserves so few macro invocations. One main reason we found is that there are usually multiple macro invocations per line, and per-line will expand all of them if any tokens in this line is changed.

RQ2

Our approach and per-file lead to no errors while several errors are caused by per-line. This is because there are quite a few macro invocations that cross multiple lines. These macros take expressions or statements as argument, which are usually too long to be included in one line.

RQ3

As discussed before, our approach may report a failure during the backward transformation. This is usually because the changes accidentally introduce a new macro invocation in the preprocessed code, where there is no way to satisfy PUTGET. However, we do not observe any such cases in our experiment. The reason is that macros usually have special names and it is not easy to collide with a macro name by copying or mutation. Note the other two approaches never report a failure, so the corresponding fields in Table 3 are left blank.

Although probably being rare in practice, theoretically our approach may report false alarms: our approach reports a failure but a correct change on the source program exists. For example, let consider the following code piece,

```
#define p (x)
plus p
```

where `plus` is the macro defined in code piece (4). After preprocessing, this code piece becomes `plus (x)`. If we change the last parenthesis into `) hello`, our approach reports a failure because first `p` will be expanded and then the expanded content forms a new macro invocation with `plus`. However, there exists a feasible change: replacing `p` with `hello p`.

0.5 Related Work

0.5.1 Bidirectional Transformation

Our work is inspired by research on bidirectional transformation. A classical scenario is the *view-update problem* [22–26](#) from database design: a view represents a database computed for a source by a query, and the problem comes when translating an update of the view back to a corresponding update on the source.

Languages have been designed to streamline the development of such applications involving transformations running bidirectionally. Notably the *lenses* framework [19, 27–35](#), covering a number of languages that provide bidirectional combinators as language constructs. A different approach is to mechanically transform existing unidirectional programs to obtain a backward counterpart, a technique known as *bidirectionalization* [15–17, 36–42](#). In the software model transformation literature, the underlying data to be transformed are usually in the form of graphs (instead of trees), and a relational (as oppose to functional) approach that specifies the bidirectional mappings between different model formats is more common [43–48](#). However, the requirement of our work goes beyond what these languages offer: in our framework, not only data, but also transformations (macros) are subject to bidirectional updates.

0.5.2 Analyzing and editing unpreprocessed C code

The C preprocessor poses a great challenge for static program analyses. The ability of producing a number of possible preprocessed variants causes a combinatorial explosion, rendering it infeasible to employ traditional tools that are designed to analyze a single variant at a time. Only until very recently, sound parsing and analyzing unpreprocessed C code is made possible through *family-based analyses* [49–51](#). Earlier tools have to resort to unsound heuristics or restrict to specific usage patterns [52–54](#).

Similarly, a lot of efforts in refactoring C code are devoted into dealing with multiple variants. Most approaches [13, 55–57](#) try to find a suitable model that represent both the C program and the preprocessor directives. A recent approach [58](#) suggests an alternative: perform refactoring on one variant and prevent the refactoring if problems may be caused

in other variants. This is based on the observation that changes on one variant seldom causes problems in other variant.

Unlike these approaches, our approach currently considers only one variant. In the future we may combine our approach with these approaches to deal with multiple variants. However, handling only one variant is already useful in many cases: (1) many programs, though with conditional compilation, do not have many variants; (2) as revealed by Overbey et al. [58](#), changes in one variant often do not cause problems in other variants.

0.5.3 Empirical studies on the C preprocessors

Over the years, there has been no shortage of academic empirical studies that are critical towards the C preprocessor [8](#), [59](#), [60](#), and replacements of CPP are proposed such as syntactical preprocessors [12](#), [61](#) and aspect-oriented programming [62–64](#) are plenty. However until present, there is no sign of any adoption of these alternatives in industry, with the C preprocessor is still being seen as the tool of the trade **Medeiros2015** .

0.6 Conclusion

Handling the C preprocessor in program-editing tools is difficult, as a result many tools either produce unsound results or give up on handling CPP entirely. In this paper we show that we can separate the concerns by using bidirectional transformations to deal with the preprocessor, so that program-editing tools may focus only on the preprocessed code, achieving a more modular design.

CPP also represents a family of transformation systems where the transformation program and the source data are bound together, for example bidirectional PHP [65](#). Existing approaches [65](#) often resort to ad-hoc treatment of the bidirectionalizing algorithm and correctness reasoning. The algorithm in this paper indicates a plausible systematic way of bidirectionalizing such systems: treating programs as data and bidirectionalize on the high-level operational semantics. We would like to see a general theory to be developed and applied to many different systems.

参考文献

- [63] B. Adams *et al.* “Can We Refactor Conditional Compilation into Aspects?” In: *Proceedings of the 8th ACM International Conference on Aspect-oriented Software Development*. Charlottesville, Virginia, USA: ACM, **2009**: 243–254. <http://doi.acm.org/10.1145/1509239.1509274>.
- [22] F. Bancilhon and N. Spyrtos. “Update Semantics of Relational Views”. *ACM Trans. Database Syst.* **1981**, 6(4): 557–575.
- [52] I. D. Baxter and M. Mehlich. “Preprocessor Conditional Removal by Simple Partial Evaluation”. In: *Proceedings of the Eighth Working Conference on Reverse Engineering (WCRE’01)*. Washington, DC, USA: IEEE Computer Society, **2001**: 281–. <http://dl.acm.org/citation.cfm?id=832308.837146>.
- [29] Ed. by G. C. Necula and P. Wadler. “Boomerang: resourceful lenses for string data”. In: *POPL*. ACM, **2008**: 407–419.
- [64] Q. Boucher *et al.* “Tag and Prune: A Pragmatic Approach to Software Product Line Implementation”. In: *Proceedings of the IEEE/ACM International Conference on Automated Software Engineering*. Antwerp, Belgium: ACM, **2010**: 333–336. <http://doi.acm.org/10.1145/1858996.1859064>.
- [25] Y. Cui, J. Widom and J. L. Wiener. “Tracing the Lineage of View Data in a Warehousing Environment”. *ACM Trans. Database Syst.* 2000-06: 179–227. <http://doi.acm.org/10.1145/357775.357777>.
- [23] U. Dayal and P. A. Bernstein. “On the Correct Translation of Update Operations on Relational Views”. *ACM Trans. Database Syst.* **1982**, 7(3): 381–416.
- [32] Ed. by J. Whittle, T. Clark and T. Kühne. “From State- to Delta-Based Bidirectional Model Transformations: The Symmetric Case”. In: *Model Driven Engineering Languages and Systems*. Springer Berlin Heidelberg, **2011**: 304–318. http://dx.doi.org/10.1007/978-3-642-24485-8_22.
- [8] M. D. Ernst, G. J. Badros and D. Notkin. “An empirical analysis of C preprocessor use”. *Software Engineering, IEEE Transactions on*, **2002**, 28(12): 1146–1170.
- [26] L. Fegaras. “Propagating updates through XML views using lineage tracing”. In: *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, 2010-03: 309–320.

- [14] B. Fluri *et al.* “Change distilling: Tree differencing for fine-grained source code change extraction”. *Software Engineering, IEEE Transactions on*, **2007**, 33(11): 725–743.
- [30] Ed. by J. Hook and P. Thiemann. “Quotient lenses”. In: *ICFP*. ACM, **2008**: 383–396.
- [19] J. N. Foster *et al.* “Combinators for Bidirectional Tree Transformations: A Linguistic Approach to the View-update Problem”. *ACM Trans. Program. Lang. Syst.* 2007-05.
- [34] Ed. by J. Gibbons. “Three Complementary Approaches to Bidirectional Programming”. In: *SSGIP*. Springer, **2010**: 1–46.
- [53] A. Garrido and R. Johnson. “Analyzing Multiple Configurations of a C Program”. In: *Proceedings of the 21st IEEE International Conference on Software Maintenance*. Washington, DC, USA: IEEE Computer Society, **2005**: 379–388. <http://dx.doi.org/10.1109/ICSM.2005.23>.
- [55] A. Garrido and R. Johnson. “Challenges of Refactoring C Programs”. In: *Proceedings of the International Workshop on Principles of Software Evolution*. Orlando, Florida: ACM, **2002**: 6–14. <http://doi.acm.org/10.1145/512035.512039>.
- [13] A. Garrido and R. Johnson. “Embracing the C preprocessor during refactoring”. *Journal of Software: Evolution and Process*, **2013**, 25(12): 1285–1304. <http://dx.doi.org/10.1002/smr.1603>.
- [50] P. Gazzillo and R. Grimm. “SuperC: Parsing All of C by Taming the Preprocessor”. In: *Proceedings of the 33rd ACM SIGPLAN Conference on Programming Language Design and Implementation*. Beijing, China: ACM, **2012**: 323–334. <http://doi.acm.org/10.1145/2254064.2254103>.
- [24] Ed. by S. Abiteboul and P. C. Kanellakis. “Foundations of Canonical Update Support for Closed Database Views”. In: *ICDT*. Springer, **1990**: 422–436.
- [47] Ed. by P. Hudak and S. Weirich. “Bidirectionalizing graph transformations”. In: *ICFP*. ACM, **2010**: 205–216.
- [48] S. Hidaka *et al.* “GRoundTram: An Integrated Framework for Developing Well-behaved Bidirectional Model Transformations”. In: *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering*. Washington, DC, USA: IEEE Computer Society, **2011**: 480–483. <http://dx.doi.org/10.1109/ASE.2011.6100104>.
- [33] M. Hofmann, B. Pierce and D. Wagner. “Edit Lenses”. In: *Proceedings of the 39th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. Philadelphia, PA, USA: ACM, **2012**: 495–508. <http://doi.acm.org/10.1145/2103656.2103715>.
- [27] Ed. by N. Heintze and P. Sestoft. “A programmable editor for developing structured documents based on bidirectional transformations”. In: *PEPM*. ACM, **2004**: 178–189.
- [20] ISO/IEC JTC 1, SC 22, WG 14. *Programming languages – C*, 2011-04.
- [49] C. Kästner *et al.* “Variability-aware Parsing in the Presence of Lexical Macros and Conditional Compilation”. In: *Proceedings of the 2011 ACM International Conference on Object Oriented Programming Systems Languages and Applications*. Portland, Oregon, USA: ACM, **2011**: 805–824. <http://doi.acm.org/10.1145/2048066.2048128>.

- [21] D. Kim *et al.* “Automatic patch generation learned from human-written patches”. In: *ICSE*, **2013**: 802–811.
- [9] E. Kohlbecker *et al.* “Hygienic macro expansion”. In: *LISP and functional programming*, **1986**: 151–161.
- [11] J. Korpela. *Using a C preprocessor as an HTML authoring tool*, **2000**.
- [2] C. Le Goues *et al.* “A systematic study of automated program repair: Fixing 55 out of 105 bugs for \$8 each”. In: *ICSE*, **2012**: 3–13.
- [1] C. Le Goues *et al.* “GenProg: A generic method for automatic software repair”. *Software Engineering, IEEE Transactions on*, **2012**, 38(1): 54–72.
- [10] B. Lee *et al.* “Marco: Safe, expressive macros for any language”. In: *ECOOP*. Springer, **2012**: 589–613.
- [5] J. Li *et al.* “SWIN: Towards Type-Safe Java Program Adaptation between APIs”. In: *PEPM*, **2015**: 91–102.
- [60] J. Liebig, C. Kästner and S. Apel. “Analyzing the Discipline of Preprocessor Annotations in 30 Million Lines of C Code”. In: *Proceedings of the Tenth International Conference on Aspect-oriented Software Development*. Porto de Galinhas, Brazil: ACM, **2011**: 191–202. <http://doi.acm.org/10.1145/1960275.1960299>.
- [51] J. Liebig *et al.* “Scalable Analysis of Variable Software”. In: *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*. Saint Petersburg, Russia: ACM, **2013**: 81–91. <http://doi.acm.org/10.1145/2491411.2491437>.
- [62] D. Lohmann *et al.* “A Quantitative Analysis of Aspects in the eCos Kernel”. In: *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006*. Leuven, Belgium: ACM, **2006**: 191–204. <http://doi.acm.org/10.1145/1217935.1217954>.
- [42] K. Matsuda and M. Wang. ““Bidirectionalization for Free” for Monomorphic Transformations”. *Science of Computer Programming*, **2014**. <http://www.sciencedirect.com/science/article/pii/S016764>
- [40] Ed. by R. Peña and T. Schrijvers. “Bidirectionalization for free with runtime recording: or, a light-weight approach to the view-update problem”. In: *PPDP*. ACM, **2013**: 297–308.
- [39] Ed. by M. Felleisen and P. Gardner. “FliPpr: A Prettier Invertible Printing System”. In: *Programming Languages and Systems*. Springer Berlin Heidelberg, **2013**: 101–120. http://dx.doi.org/10.1007/978-3-642-37036-6_6.
- [18] Ed. by A. D. Gordon. “A Grammar-Based Approach to Invertible Programs”. In: *ESOP*. Springer, **2010**: 448–467.
- [15] Ed. by R. Hinze and N. Ramsey. “Bidirectionalization transformation based on automatic derivation of view complement functions”. In: *ICFP*. ACM, **2007**: 47–58.

- [12] B. McCloskey and E. Brewer. “ASTEC: A New Approach to Refactoring C”. In: *ESEC/FSE-13*, **2005**: 21–30.
- [7] N. Meng, M. Kim and K. S. McKinley. “Systematic Editing: Generating Program Transformations from an Example”. In: *PLDI*, **2011**: 329–342.
- [28] Ed. by W.-N. Chin. “An Algebraic Approach to Bi-directional Updating”. In: *APLAS*. Springer, **2004**: 2–20.
- [4] H. D. T. Nguyen *et al.* “SemFix: Program repair via semantic analysis”. In: *ICSE*, **2013**: 772–781.
- [43] I. (Object Management Group. *Meta Object Facility (MOF) 2.0 Query/View/Transformation Specification*.
- [58] J. L. Overbey, F. Behrang and M. Hafiz. “A Foundation for Refactoring C with Macros”. In: *Proceedings of the 22Nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. Hong Kong, China: ACM, **2014**: 75–85. <http://doi.acm.org/10.1145/2635868.2635908>.
- [54] Y. Padioleau. “Parsing C/C++ Code Without Pre-processing”. In: *Proceedings of the 18th International Conference on Compiler Construction: Held As Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2009*. York, UK: Springer-Verlag, **2009**: 109–125. http://dx.doi.org/10.1007/978-3-642-00722-4_9.
- [6] Y. Padioleau *et al.* “Semantic Patches for Documenting and Automating Collateral Evolutions in Linux Device Drivers”. In: *PLOS*, **2006**.
- [3] Y. Qi *et al.* “The strength of random search on automated program repair”. In: *ICSE*, **2014**: 254–265.
- [35] R. Rajkumar *et al.* *Lenses for Web Data*, **2013**.
- [45] Ed. by E. Mayr, G. Schmidt and G. Tinhofer. “Specification of graph translators with triple graph grammars”. In: *Graph-Theoretic Concepts in Computer Science*. Springer Berlin Heidelberg, **1995**: 151–163. http://dx.doi.org/10.1007/3-540-59071-4_45.
- [46] Ed. by H. Ehrig *et al.* “15 Years of Triple Graph Grammars”. In: *Graph Transformations*. Springer Berlin Heidelberg, **2008**: 411–425. http://dx.doi.org/10.1007/978-3-540-87405-8_28.
- [59] H. Spencer and G. Collyer. *#ifdef Considered Harmful, or Portability Experience With C News*, **1992**.
- [57] D. Spinellis. “Global Analysis and Transformations in Preprocessed Languages”. *IEEE Trans. Softw. Eng.* 2003-11: 1019–1030. <http://dx.doi.org/10.1109/TSE.2003.1245303>.
- [44] P. Stevens. “Bidirectional model transformations in QVT: semantic issues and open questions”. *Software & Systems Modeling*, **2010**, 9(1): 7–20. <http://dx.doi.org/10.1007/s10270-008-0109-9>.

- [56] M. Vittek. “Refactoring browser with preprocessor”. In: *Software Maintenance and Reengineering, 2003. Proceedings. Seventh European Conference on*, 2003-03: 101–110.
- [16] Ed. by Z. Shao and B. C. Pierce. “Bidirectionalization for free! (Pearl)”. In: *POPL*. ACM, **2009**: 165–176.
- [17] J. Voigtländer *et al.* “Combining syntactic and semantic bidirectionalization”. In: *ICFP*, **2010**: 181–192.
- [37] J. Voigtländer *et al.* “Enhancing semantic bidirectionalization via shape bidirectionalizer plugins”. *J. Funct. Program.* **2013**, 23(5): 515–551.
- [36] Ed. by M. M. T. Chakravarty, Z. Hu and O. Danvy. “Incremental updates for efficient bidirectional transformations”. In: *ICFP*. ACM, **2011**: 392–403.
- [41] Ed. by W.-N. Chin and J. Hage. “Semantic bidirectionalization revisited”. In: *PEPM*. ACM, **2014**: 51–62.
- [31] Ed. by C. Bolduc, J. Desharnais and B. Ktari. “Gradual Refinement: Blending Pattern Matching with Data Abstraction”. In: *MPC*. Springer, **2010**: 397–425.
- [38] M. Wang *et al.* “Refactoring pattern matching”. *Sci. Comput. Program.* **2013**, 78(11): 2216–2242.
- [65] X. Wang *et al.* “Automating presentation changes in dynamic web applications via collaborative hybrid analysis”. In: *FSE*, **2012**: 16.
- [61] D. Weise and R. Crew. “Programmable Syntax Macros”. In: *Proceedings of the ACM SIGPLAN 1993 Conference on Programming Language Design and Implementation*. Albuquerque, New Mexico, USA: ACM, **1993**: 156–165. <http://doi.acm.org/10.1145/155090.155105>.

附录 A 附件

pkuthss 文档模版最常见问题:

在最终打印和提交论文之前, 请将 *pkuthss* 文档类选项中的 `colorlinks` 改为 `nocolorlinks`, 因为图书馆要求电子版论文的目录必须为黑色, 且某些教务要求打印版论文的文字部分为纯黑色而非灰度打印。

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记: `test-en`, `[test-zh]`、`[test-en, test-zh]`。

若要避免章末空白页, 请在调用 *pkuthss* 文档类时加入 `openany` 选项。

如果编译时不出参考文献, 请参考 `texdoc pkuthss` “问题及其解决”一章“其它可能存在的问题”一节中关于 `biber` 的说明。

致谢

pkuthss 文档模版最常见问题:

在最终打印和提交论文之前, 请将 *pkuthss* 文档类选项中的 **colorlinks** 改为 **nocolorlinks**, 因为图书馆要求电子版论文的目录必须为黑色, 且某些教务要求打印版论文的文字部分为纯黑色而非灰度打印。

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记: **test-en**, **[test-zh]**、**[test-en, test-zh]**。

若要避免章末空白页, 请在调用 *pkuthss* 文档类时加入 **openany** 选项。

如果编译时不出参考文献, 请参考 `texdoc pkuthss` “问题及其解决”一章“其它可能存在的问题”一节中关于 `biber` 的说明。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校在 ☐ 一年 / ☐ 两年 / ☐ 三年以后在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名： 日期： 年 月 日