Code Duplication

Yegor Bugayenko

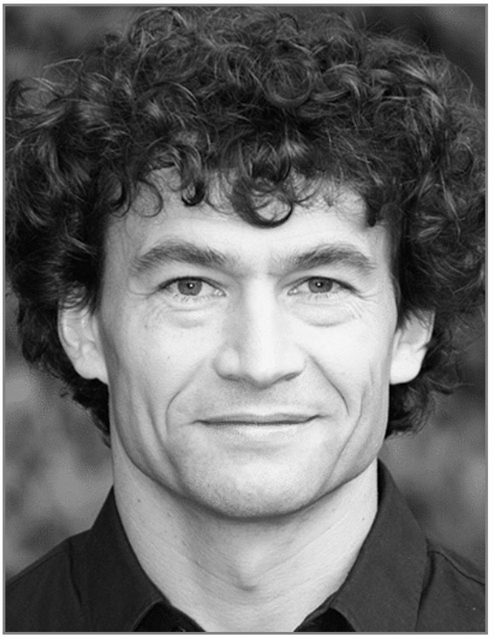
Lecture #11 out of 24
80 minutes

+2pt

"If a bug is identified in one segment of code, all the similar segments need to be checked for the same bug. Consequently, code duplication may lead to bug propagation that significantly affects the maintenance cost."

— *A Systematic Review on Code Clone Detection*, Qurat Ul Ain, Wasi Haider Butt, Muhammad Waseem Anwar, Farooque Azam, Bilal Maqbool, IEEE Access, 2019

+2pt



"The problem with code cloning is that errors in the original must be fixed in every copy. Other kinds of maintenance changes, for instance, extensions or adaptations, must be applied multiple times, too. Yet, it is usually not documented where code was copied."

— *Comparison and Evaluation of Clone Detection Tools*, Stefan Bellon, Rainer Koschke, Giuliano Antoniol, Jens Krinke, Ettore Merlo, IEEE Transactions on Software Engineering, 2007

Motivating Example (part I)

Before (wrong):

```
printf("Hi,%s!",getName(42));
printf("Hi,%s!",getName(7));
printf("Hi,%s!",getName(55));
```

After (better):

```
sayHello(42);
sayHello(7);
sayHello(55);

void sayHello(int id) {
  var n = getName(id);
  printf("Hi,%s!",n);
}
```

Motivating Example (part II)

Before (still not ideal):

```
sayHello(42);
sayHello(7);
sayHello(55);

void sayHello(int id) {
  var n = getName(id);
  printf("Hi,%s!",n);
}
```

After (perfect):

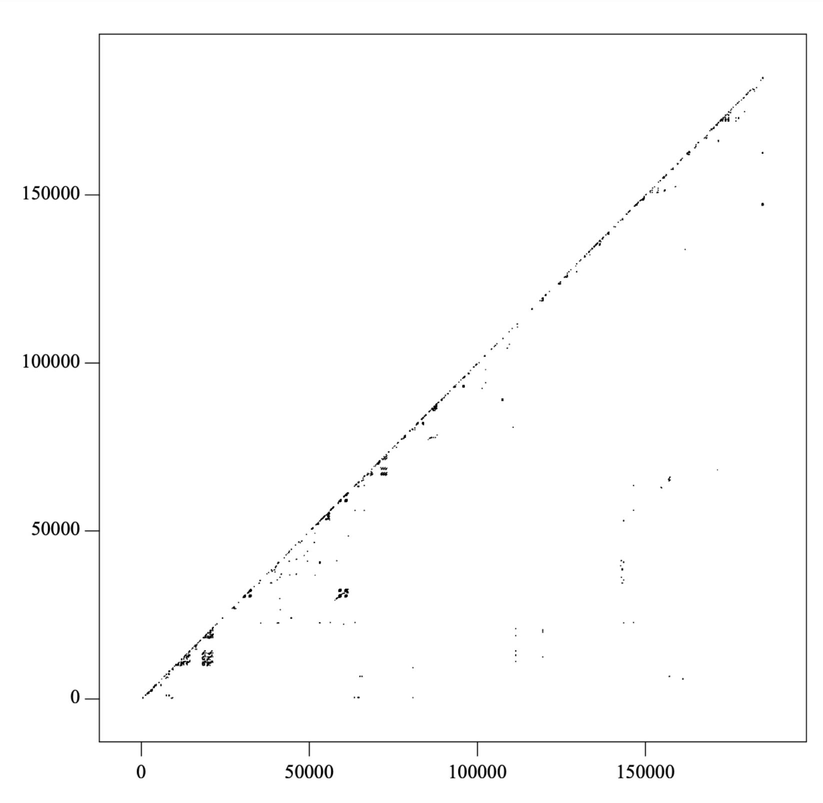
```
var users = [42,7,55];
for (id : users) {
  sayHello(id);
}

void sayHello(int id) {
  var n = getName(id);
  printf("Hi,%s!",n);
}
```

+2pt

> “Two lines of code are considered to be identical if they contain the same sequence of characters after removing comments and white space; the semantics of the program statements are not analyzed.”
>
> — *A Program for Identifying Duplicated Code*, Brenda S. Baker, Computing Science and Statistics, 1993

Up to 38% of lines are involved in duplicates



The plots are dense near the main diagonal, implying that most copies tend to occur fairly locally, e.g. within the same file or module.

However, certain line segments occur away from the main diagonal; it would be interesting to investigate why the corresponding sections of code are duplicated.

Don’t Repeat Yourself (DRY)

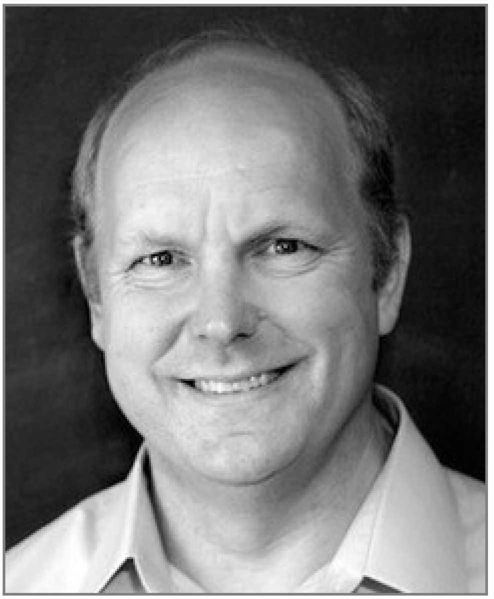
+2pt

“Every piece of knowledge must have a single, unambiguous, authoritative representation within a system.”

— *The Pragmatic Programmer: From Journeyman to Master*, Andrew Hunt and David Thomas, Addison-Wesley, 1999

The Rule of Three

+2pt



"The first time you do something, you just do it. The second time you do something similar, you wince at the duplication, but you do the duplicate thing anyway. The third time you do something similar, you refactor."

— *Refactoring*, Martin Fowler and Kent Beck, Addison-Wesley, 1999

+2pt

"We identified and analyzed 26 Code Clone Detection (CCD) tools, i.e., 13 existing and 13 proposed/developed. Moreover, 62 open-source subject systems whose source code is utilized for the CCD are presented."

— *A Systematic Review on Code Clone Detection*, Ain, Qurat Ul, Wasi Haider Butt, Muhammad Waseem Anwar, Farooque Azam, Bilal Maqbool, IEEE Access, 2019

Type-1: Exact Clone

Original:

```
printf("Hi,%s\n",name(42));
```

Clone:

```
// Here we print a message
// to the console for a user
printf(
  "Hi,%s\n",
  name(42)
);
```

Identical code segments except for changes in comments, layouts and whitespaces.

Type-2: Parameterized Clone

Original:

```
var n = name(42);
printf("Hi,%s\n",n);
```

Clone:

```
String name = name(42);
printf("Hi,%s\n",name);
```

Code segments which are syntactically or structurally similar other than changes in comments, identifiers, types, literals, and layouts.

Type-3: Gapped Clone

Original:

```
printf("Hi,%s\n",name(42));
```

Clone:

```
var msg = "Hi,%s\n";
var n = name(42);
printf(msg,n);
```

Copied pieces with further modification such as addition or removal of statements and changes in whitespaces, identifiers, layouts, comments, and types but outcomes are similar.

Type-4: Semantic Clone

Original:

```
printf("Hi,%s\n",name(42));
```

Clone:

```
var s = sprintf(
  "Hi,%s\n",
  name(42));
print(s);
```

More than one code segments that are functionally similar but implemented by different syntactic variants.

These tools can help detecting duplicate code:

1. IntelliJ IDEA by JetBrains

2. Copy/Paste Detector (CPD) by PMD for Java

Read this:

A Program for Identifying Duplicated Code, Brenda S. Baker, Computing Science and Statistics, 1993

A Systematic Review on Code Clone Detection, Qurat Ul Ain, Wasi Haider Butt, Muhammad Waseem Anwar, Farooque Azam, Bilal Maqbool, IEEE Access, 2019

A Systematic Review on Code Clone Detection, Qurat Ul Ain, Wasi Haider Butt, Muhammad Waseem Anwar, Farooque Azam, Bilal Maqbool, IEEE Access, 2019

Comparison and Evaluation of Clone Detection Tools, Stefan Bellon, Rainer Koschke, Giuliano Antoniol, Jens Krinke, Ettore Merlo, IEEE Transactions on Software Engineering, 2007

Refactoring, Martin Fowler and Kent Beck, Addison-Wesley, 1999