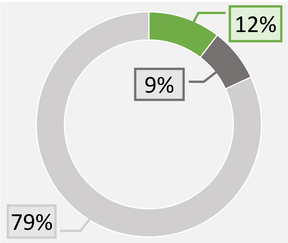
UC Irvine : Crista **Lopes**, Pedro **Martins**, Vaibhav **Saini**, Di **Yang**, Hitesh **Sajnani** — **CTU Prague** : Petr **Maj**, Jakub **Zitny** — **Northeastern** : Jan **Vitek**

DéjàVu

A Map of Code Duplicates on GitHub

Introduction



If imitation is flattery, then the Internet is surprisingly full of praise. Specifically, the code programmers check into large scale software repositories is riddled with duplicates. Our paper involves the analysis of 4.5M non-fork projects on GitHub; 428M files in Java, C++, Python, and JavaScript. We found only 85M unique files. In other words, 79% of GitHub consists of duplicates. There is variability between languages. JavaScript has the highest rate, only 4% of the files are distinct. Java has the least, 60% of files are distinct.

Similarity Levels

```
// This is a function
void foobar(int foo, int bar) {
    int baz = foo + bar;
    ++baz;
    if (baz > 42)
        return foo;
    else
        return bar;
}
```

FILE 1
```
void : 1
foobar : 1
int : 3
foo : 3
bar : 3
baz : 3
if : 1
42 : 1
return : 2
else : 1
```

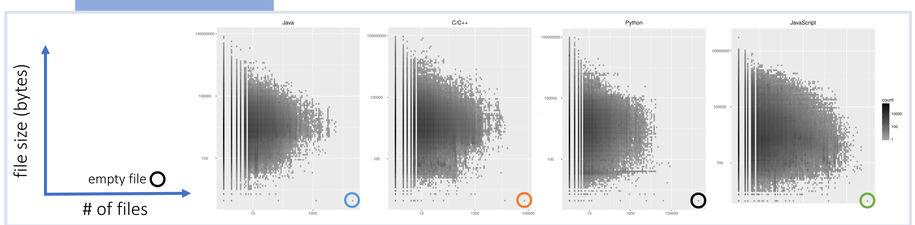
FILE 2
```
for : 1
while : 1
xy : 3
78 : 3
89 : 3
1 : 1
3 : 1
```

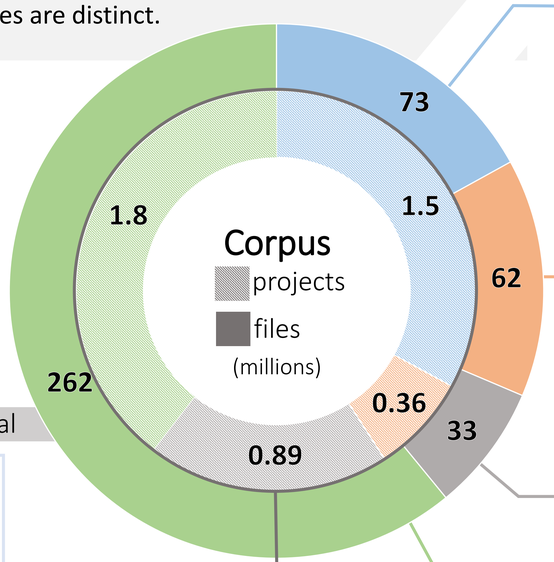
FILE 3
```
void : 1
foobar : 1
int : 2
double : 1
foo : 3
bar : 3
baz : 3
if : 1
43 : 1
return : 2
else : 1
```

FILE 4
```
double : 3
for : 2
while : 1
xy : 3
78 : 3
89 : 3
1 : 1
3 : 1
```

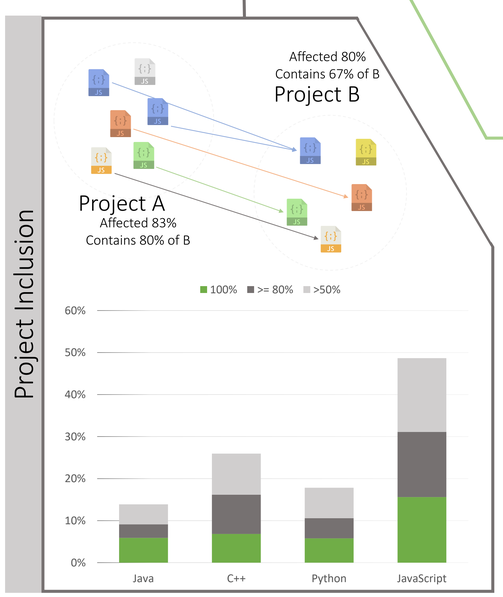
SourcererCC

File Hash | Token Hash | Near-identical

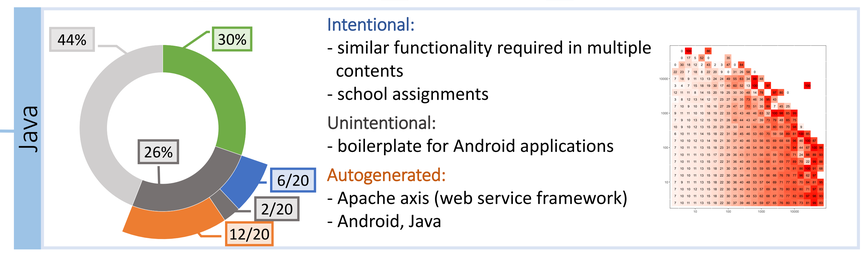




Project Inclusion



Java

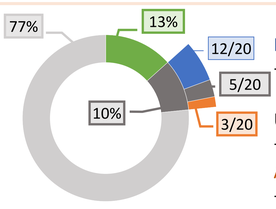


Intentional:
- similar functionality required in multiple contents
- school assignments

Unintentional:
- boilerplate for Android applications

Autogenerated:
- Apache axis (web service framework)
- Android, Java

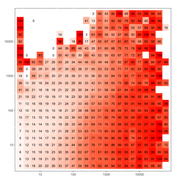
C/C++



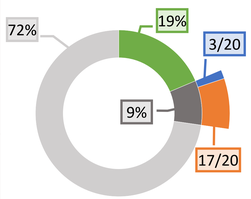
Intentional:
- slight adaptation of algorithms for different interfaces

Unintentional:
- different library versions

Autogenerated:
- Qt Meta-Object Compiler

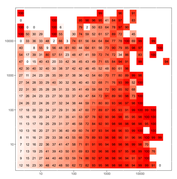


Python

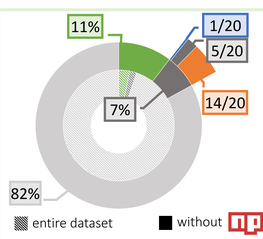


Intentional:
- unittests
- database schemes

Autogenerated:
- all 17 examples are Django



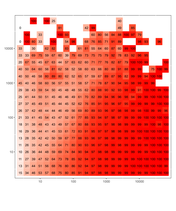
JavaScript



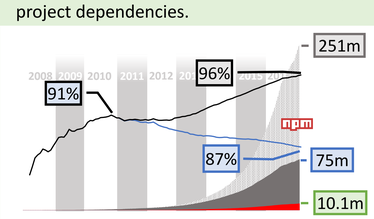
Intentional:
- test with only changed name

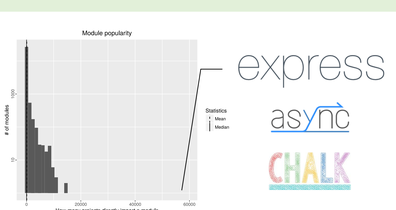
Unintentional:
- different library versions (jQuery)

Autogenerated:
- angular.js with Yeoman
- expresss.js, grunt.js



In JavaScript, over 70% of the identical duplicates of files come from accidental inclusion of node.js project dependencies.

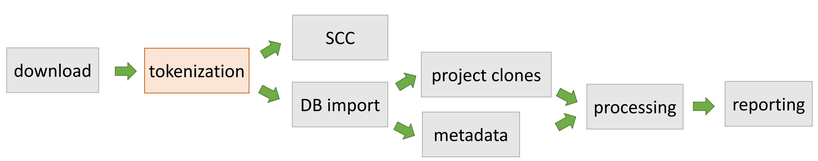




express
async
CHALK

Artifact

All our data and results are available for download. Furthermore our entire pipeline, from downloading the projects to producing the clone maps is entirely automated in the form of R notebooks.



References

http://mondego.ics.uci.edu/projects/dejavu - project website
https://github.com/PRL-PRG/dejavu-artifact - artifact
http://dejavu.ics.uci.edu – web service