An Empirical Assessment of Bellon's Clone Benchmark

Alan Charpentier[1], Jean-Rémy Falleri[1], David Lo[2] and Laurent Réveillère[1]

[1]University of Bordeaux, France
[2] Singapore Management University, Singapore

EASE'15 - Nanjing, China

All Starts with Clones

Clones

- A clone refers to fragments of code that are *similar* or *identical*.
- Cloning can complicate code maintenance and evolution.

Clone detectors

- Scan code base for potential clones.
- Find incorrect clones (false positives).
- Miss true clones (false negatives).

Clone detectors enhancement

- Clone benchmarks are used to compare and assess clone detectors' results.

Bellon Benchmark[1]

The construction

- Six researchers provided clones to Bellon.
- Bellon examined 2 percent of all $325,935$ submitted clones.
- Bellon built a *reference corpus* by retaining only clones he judged as true positives.
- Clone definition used by Bellon:
  - *A clone is a triple* (f_1, f_2, t) *where* f_1 *and* f_2 *are two similar code fragments and* t *is the associated type of similarity.*
  - Clones are pairs of code fragments that could be replaced by function calls.
  - Code fragments of a clone must contain at least six lines of code.

[1] S. Bellon, R. Koschke, G. Antoniol, J. Krinke and E. Merlo. Comparison and evaluation of clone detection tools. *Software Engineering, IEEE Transactions on*, 33(9):577-591, Sept 2007.

Bellon Benchmark

The use

Two-step process

- Computation of a mapping between candidate clones and reference clones.
- Computation of the recall and precision effectiveness measures.
  - $$Recall(P, T, \tau) = \frac{|DetectedRefs(P, T, \tau)|}{|Refs(P, \tau)|}$$
  - $$Precision(P, T, \tau) = \frac{|DetectedRefs(P, T, \tau)|}{|Cands(P, \tau)|}$$

Agreement on the Reference Clones

- Clones are very subjective.
- Bellon built alone the reference corpus.
- Bellon was not an expert of the projects containing the clones.

- Can we trust the references clones?
- To what extent other persons would agree that the reference clones are indeed true clones?

Research Questions

- **RQ1:** Can researchers trust the clones from Bellon's reference corpus?
- **RQ2:** Are the effectiveness measures computed using Bellon's benchmark reliable?
- **RQ3:** Are there some characteristics of reference clones that make them more trustable?

Experiment

Overview

A reference clone can be trusted if anyone that is presented the clone won't doubt that it is a true clone.

- Selection of a subset of the reference clones.
- Presentation of these clones to additional persons.
- Gathering their opinions on the clones.

Experiment

Participants

18 students participate to the experiment.

- Four are graduate.
  - Three from Singapore.
  - One from France.
- Fourteen undergraduate.
  - All in the last year of their degree in France.
- All have been trained in Java and C programming language.

Experiment

Clone selection

Requirements

- 120 clones per participant.
- 2 participants per clone.

Selection

- $1,080$ clones randomly selected from the $4,096$ clones of the reference corpus.

Distribution

- The $1,080$ clones are randomly split in nine groups of 120 clones.
- Each group is examined by a pair of two randomly drawn students.

Experiment

Opinion collection

Preparation

- Clone definition used by Bellon was presented to participants.
- Participants do not know that the clones they are rating were judged as true clones.

Collection

- Participants had a time slot of 3 hours to make the experiment.
- The answers are collected through a web site.
- Participants are asked to rate *yes* for clones they deem as true clones, and *no* for false clones.

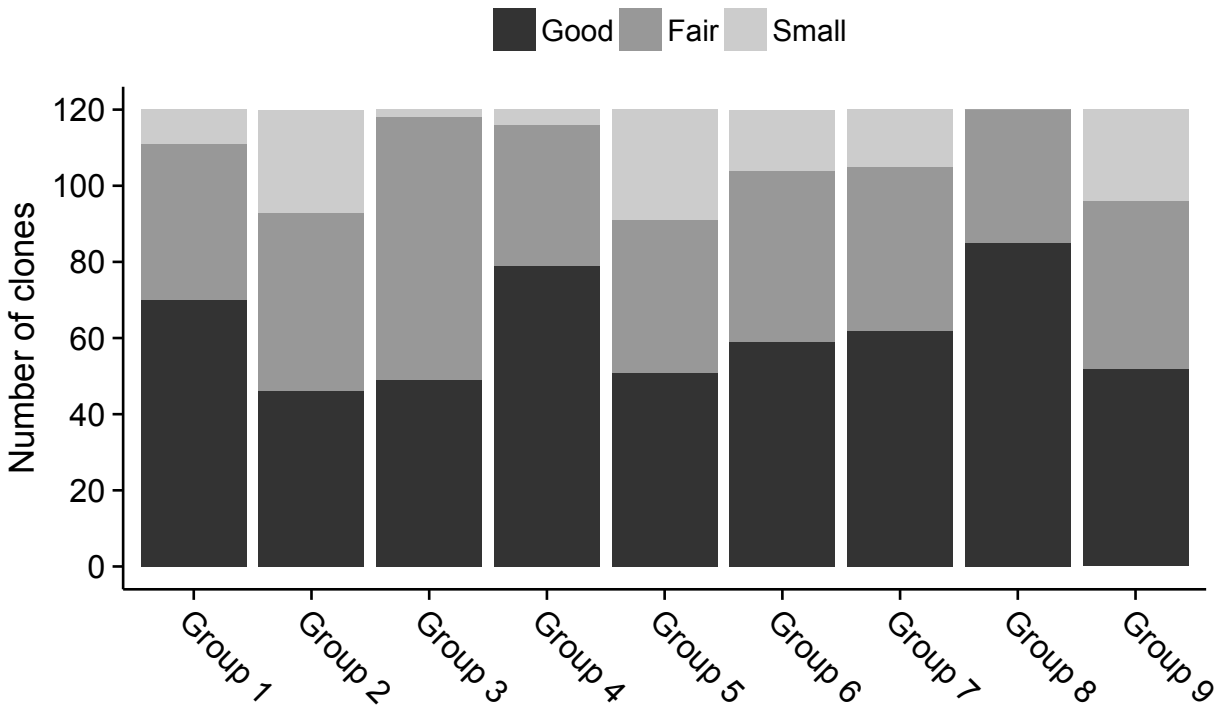
Results

- We obtain $1,080$ clones with 3 opinions each.
- We affect a trust level to each clone.
  - Thee positive opinions: *good* trust level
  - Two positive opinions: *fair* trust level
  - One positive opinion: *small* trust level
- We define the following ordinal scale: $\textit{small} < \textit{fair} < \textit{good}$

Trust Level of Reference Clones

Our sample of 1,080 clones

- About half of the clones have less than a *good* trust level.
- Between ten to fifteen percent of the clones have only a *small* trust level.



Trust Level of Reference Clones

Estimation in the reference corpus

- We use bootstrapping on our sample of $1,080$ clones.
- We compute 95% intervals for these ratios.

	Confidence interval	
Trust level	**Lower bound**	**Higher bound**
Small	0.10	0.14
Fair	0.34	0.40
Good	0.48	0.54

- There is a significant number of reference clones that are debatable.

Trust Level and Effectiveness Measures

Two scenarios

- Clones having at least a *fair* trust level.
- Clones having a *good* trust level.

Two new reference corpora

- F containing 954 clones
- G containing 553 clones

Trust Level and Effectiveness Measures

Worst-case approach to assess if the trust level of Bellon's reference corpus can modify precision and recall values.

t detects exactly clones of	**Precision and recall derived from**		
	B	F	G
B		$p = 0.88$	$p = 0.51$
F	$r = 0.88$		
G	$r = 0.51$		

- When requiring a *good* trust level, a precision and recall decrease of up to 0.49 is possible.
- When requiring only a *fair* trust level, a precision and recall decrease of up to 0.12 is possible.

Trust Level and Clone Characteristics

Three clone characteristics:

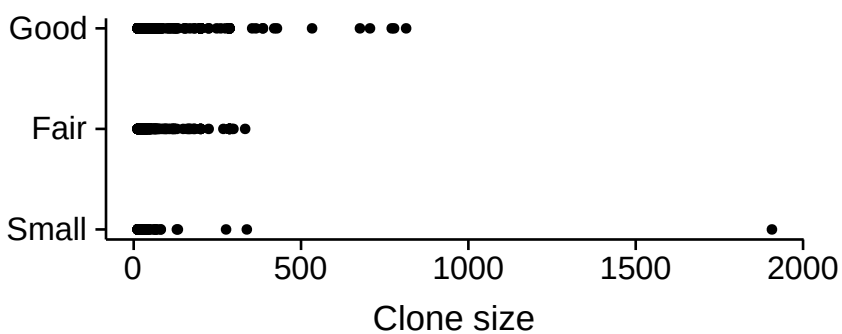
- Type
- Size
- Language

Trust Level and Clone Characteristics

Clone size

H_0: there is no correlation between a clone size and its trust level.

H_a: the larger a clone is, the greater its trust level is.



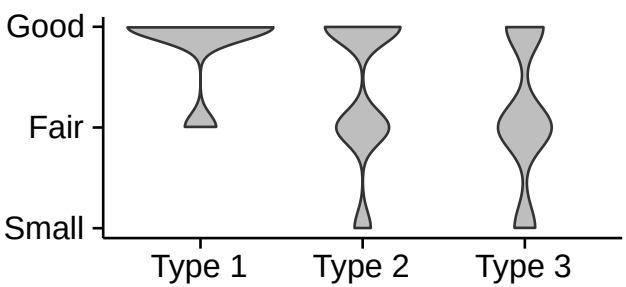
- Spearman's $\rho = 0.06$ (p-value $= 0.03$)
- 95% confidence interval: $0 \leq \rho \leq 0.12$
- Weak effect size

Trust Level and Clone Characteristics

Clone type

H_0: *there is no correlation between a clone type and its trust level.*

H_a: *the bigger the type of a clone is, the lesser its trust level is.*



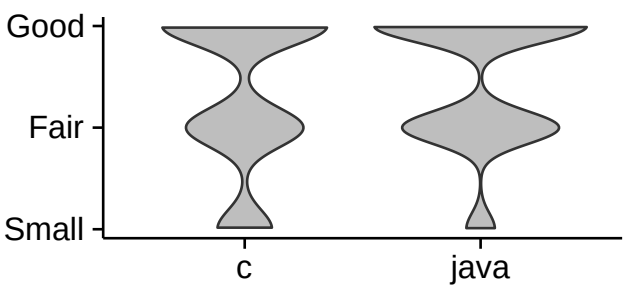
- Spearman's $\rho = -0.26$ (p-value $= 0$)
- 95% confidence interval: $-0.31 \leq \rho \leq -0.21$
- Moderate and negative effect size

Trust Level and Clone Characteristics

Programming language

H_0: the programming language has no impact on the trust level.

H_a: the programming language has an impact on the trust level.



- Mann-Whitney U test = 133672.5 (p-value = 0.01)
- Cliff's delta $d = -0.08$.
- Confidence interval: $-0.14 \leq d \leq -0.02$
- Negligible effect size

Conclusions and Future Work

Conclusions

- A significant number of reference clones are debatable.
- Precision and recall can be significantly modified by the trust level of reference clones.
- The type of a clone is the only characteristic having a significant effect size.

Future Work

- Experts to help construction of clone benchmarks.
- More precise definition to ease clone categorization.