RQ1: Can we quantify interest of TD at the functional level? How much is the interest?

Yasutaka Kamei

Feb 4th, 2016

Data Overview

```
data <- read.csv("/Users/kamei/Research/techdebt/msr16_td_interest/datasets/CSV/technical_debt_summary.c

# the number of technical debt at the method-level
nrow(data)
```

```
## [1] 837
```

Observation

- We find 837 technical debt at the method-level.
- Note that we now mix all three projects into one dataset.

How many technical debt can we map between a metrics file and Everton's summary file?

*_v1 means the version that introduces technical debt and *_v2 means the last version that technical debt was found.

```
# the number of technical debt that cannot be mapped between a metrics file
#and Everton's summary file
data.only_num <- data[,c("version_name", "CountInput_v1", "CountInput_v2",
    "CountOutput_v1", "CountOutput_v2", "CountLine_v1", "CountLine_v2",
    "Cyclomatic_v1", "Cyclomatic_v2", "MaxNesting_v1", "MaxNesting_v2")]
apply(data.only_num[,1:3], 2, function(x){sum(x == -1) })
```

```
##  version_name CountInput_v1 CountInput_v2
##             8           171           101
```

```
# the number of technical debt that have metrics in both versions
# of introduction and last_found
a <- data[(data[, "CountInput_v1"] != -1 & data[, "CountInput_v2"] != -1), ]
nrow(a)
```

```
## [1] 608
```

Observation

- We miss 171 technical debt in v1 and 101 in v2. 608 technical debt has metrics in both versions of introduction and last_found.
- We need to discuss how to solve such missed technical debt.

How much is the interest?

For 608 technical debt, we measure the interest by substracting the metric value of v2 - the metric value v1. We use 5 metrics as interest.

CountInput (fanin)

```
# interest of CountInput (fanin)
interest <- a[,"CountInput_v2"] - a[,"CountInput_v1"]
summary(interest)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -21.0000   0.0000   0.0000   0.5296   0.0000  46.0000
```

```
print( c(sum(interest==0), sum(interest > 0), sum(interest < 0)) )
```

```
## [1] 395 147  66
```

```
summary(subset(interest, interest !=0))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -21.000  -1.000   1.000   1.512   2.000  46.000
```

CountOutput (fanout)

```
# interest of CountOutput (fanout)
interest <- a[,"CountOutput_v2"] - a[,"CountOutput_v1"]
summary(interest)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -34.0000   0.0000   0.0000   0.2155   0.0000  29.0000
```

```
print( c(sum(interest==0), sum(interest > 0), sum(interest < 0)) )
```

```
## [1] 365 147  96
```

```
summary(subset(interest, interest !=0))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -34.0000  -1.0000   1.0000   0.5391   3.0000  29.0000
```

LOC

```r
# interest of LOC
interest <- a[,"CountLine_v2"] - a[,"CountLine_v1"]
summary(interest)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## -346.0000    0.0000    0.0000    0.3454    1.0000  204.0000
```

```r
print( c(sum(interest==0), sum(interest > 0), sum(interest < 0)) )
```

```
## [1] 329 171 108
```

```r
summary(subset(interest, interest !=0))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## -346.0000   -2.0000    2.0000    0.7527    6.0000  204.0000
```

Complexity

```r
# interest of Complexity
interest <- a[,"Cyclomatic_v2"] - a[,"Cyclomatic_v1"]
summary(interest)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -65.0000   0.0000   0.0000   0.3454   0.0000  66.0000
```

```r
print( c(sum(interest==0), sum(interest > 0), sum(interest < 0)) )
```

```
## [1] 448 102  58
```

```r
summary(subset(interest, interest !=0))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -65.000  -1.000   1.000   1.312   2.000  66.000
```

Max Nesting

```r
# interest of Max Nesting
interest <- a[,"MaxNesting_v2"] - a[,"MaxNesting_v1"]
summary(interest)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -5.00000  0.00000  0.00000 -0.02632  0.00000  3.00000
```

```r
print( c(sum(interest==0), sum(interest > 0), sum(interest < 0)) )
```

```
## [1] 529  38  41
```

```r
summary(subset(interest, interest !=0))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -5.0000 -1.0000 -1.0000 -0.2025  1.0000  3.0000
```

Observation

- Regardless of types of metrics, 65% of technical debt has 0 interest. So the median is 0.
- If we focus on technical debt that has more than 0 interest
  - the number of technical debt that have positive interest is more than negative interest one
  - the median is around 1-2.

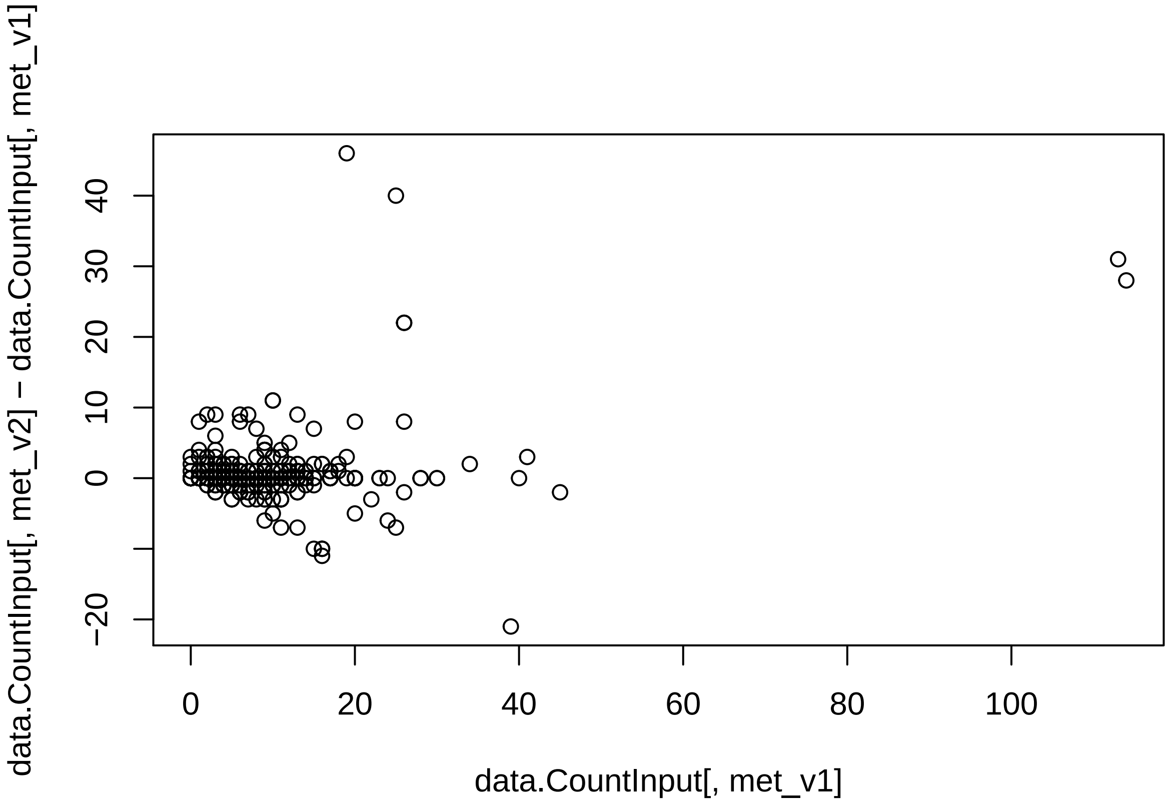
```r
# the top interest tech debt?
met_v1 <- "CountInput_v1"
met_v2 <- "CountInput_v2"

data.CountInput <- data[,c("Method_Signature", met_v1, met_v2)]
data.CountInput <- data.CountInput[(data.CountInput[, met_v1] != -1 & data.CountInput[, met_v2] != -1),
idx <- order(data.CountInput[,met_v2] - data.CountInput[,met_v1], decreasing = T)
head(data.CountInput[idx,])
```

```
##                                                             Method_Signature
## 797                                                              getObject()
## 805                              newIOErrorFromException(java.io.IOException)
## 419                                                 inspect(org.jruby.ast.Node)
## 227                                                 inspect(org.jruby.ast.Node)
## 816 format(java.util.Date, java.lang.StringBuffer, java.text.FieldPosition)
## 817 format(java.util.Date, java.lang.StringBuffer, java.text.FieldPosition)
##     CountInput_v1 CountInput_v2
## 797            19            65
## 805            25            65
## 419           113           144
## 227           114           142
## 816            26            48
## 817            26            48
```

Observation

- [Discuss] if technical debt has same version and same method siguniture, should we remove one of them?

```r
# Hypothesis: Is the method that has large fanin likely to have large interest?
#data.CountInput = subset(data.CountInput, data.CountInput$CountInput_v1 - data.CountInput$CountInput_v
plot(data.CountInput[,met_v1], data.CountInput[,met_v2] - data.CountInput[,met_v1])
```



```r
cor(data.CountInput[,met_v1], data.CountInput[,met_v2] - data.CountInput[,met_v1], method = "spearman")
```

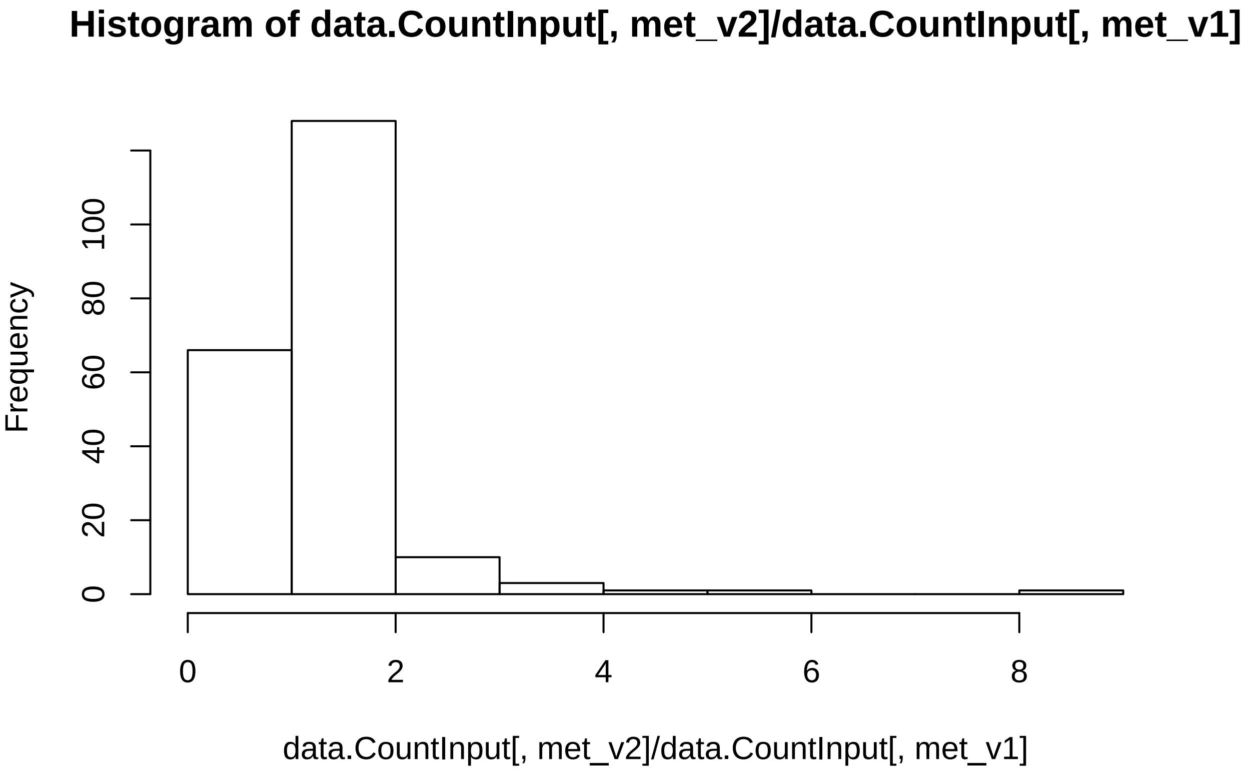
```
## [1] 0.01948702
```

Observation

- No…

[Emad] I think one thing to measure is the metric value in v-1/metric value in v-2. Of course we can only do this for non-zero differences.

```r
# Hypothesis: Is the method that has large fanin likely to have large interest?
data.CountInput = subset(data.CountInput, data.CountInput$CountInput_v1 - data.CountInput$CountInput_v2
hist(data.CountInput[,met_v2]/data.CountInput[,met_v1])
```



Observation

- There are several technical debt of which the ratio is more than 2. This means that the dependency of technical debt becomes double before being removed.