RQ Answers

omitted

2024-04-05

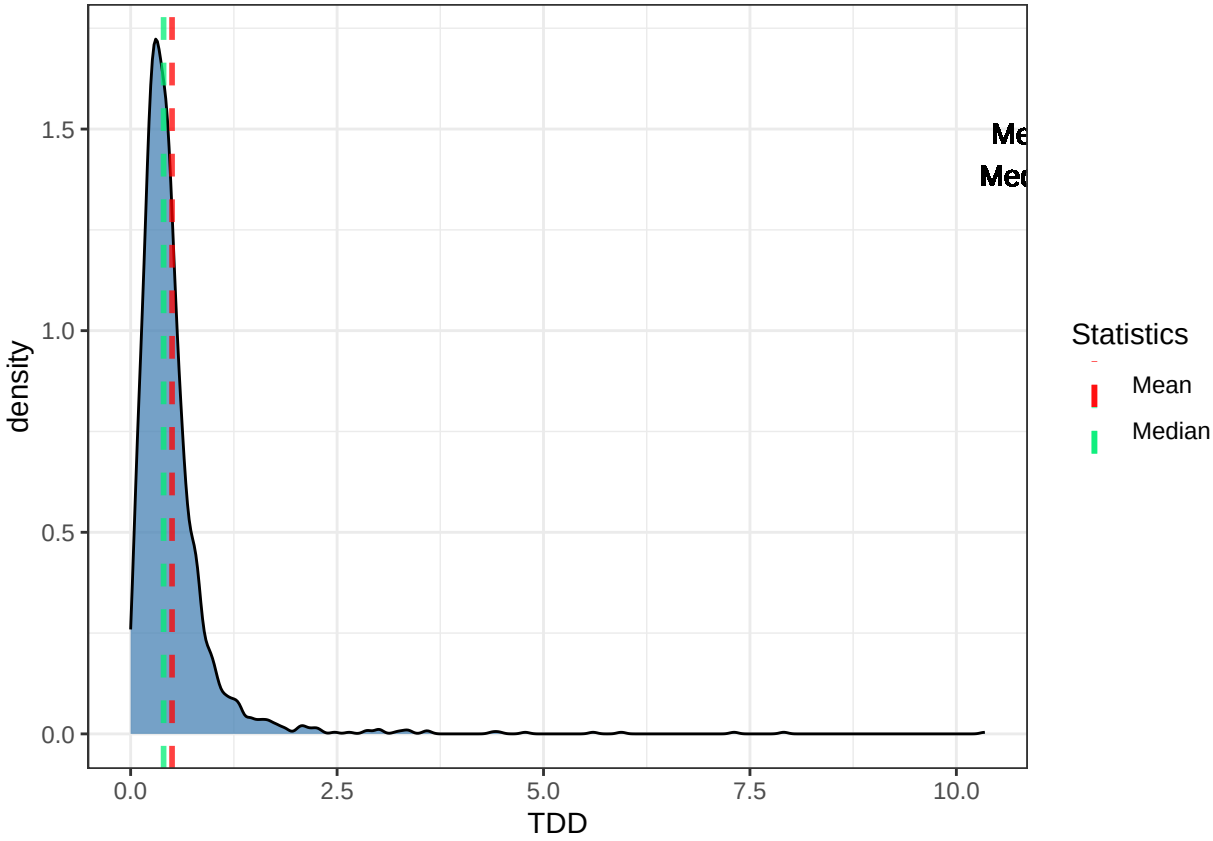
Data overview

```
glimpse(all_data)
```

```
## Rows: 247,365
## Columns: 11
## $ repo                <chr> "accumulo", "accumulo", "accumulo", "accumulo", "a~
## $ pr_number           <int> 1302, 1302, 1302, 1302, 1302, 1302, 1302, 1302, 13~
## $ rule                <chr> "java:S1192", "java:S3776", "java:S1319", "java:S1~
## $ severity            <chr> "CRITICAL", "CRITICAL", "MINOR", "CRITICAL", "CRIT~
## $ file                <chr> "core/src/main/java/org/apache/accumulo/core/metad~
## $ type                <chr> "CODE_SMELL", "CODE_SMELL", "CODE_SMELL", "CODE_SM~
## $ status              <chr> "OPEN", "OPEN", "OPEN", "OPEN", "CLOSED", "OPEN", ~
## $ debt                <int> 8, 6, 10, 5, 6, 2, 5, 15, 30, 109, 30, 30, 5, 5, 5~
## $ ncloc_affected_file <int> 347, 347, 347, 347, 347, 347, 347, 347, 252, 252, ~
## $ complexity          <int> 68, 68, 68, 68, 68, 68, 68, 68, 86, 86, 86, 86, 86~
## $ origin              <chr> "NEW", "NEW", "NEW", "NEW", "PRE-EXISTING", "PRE-E~
```

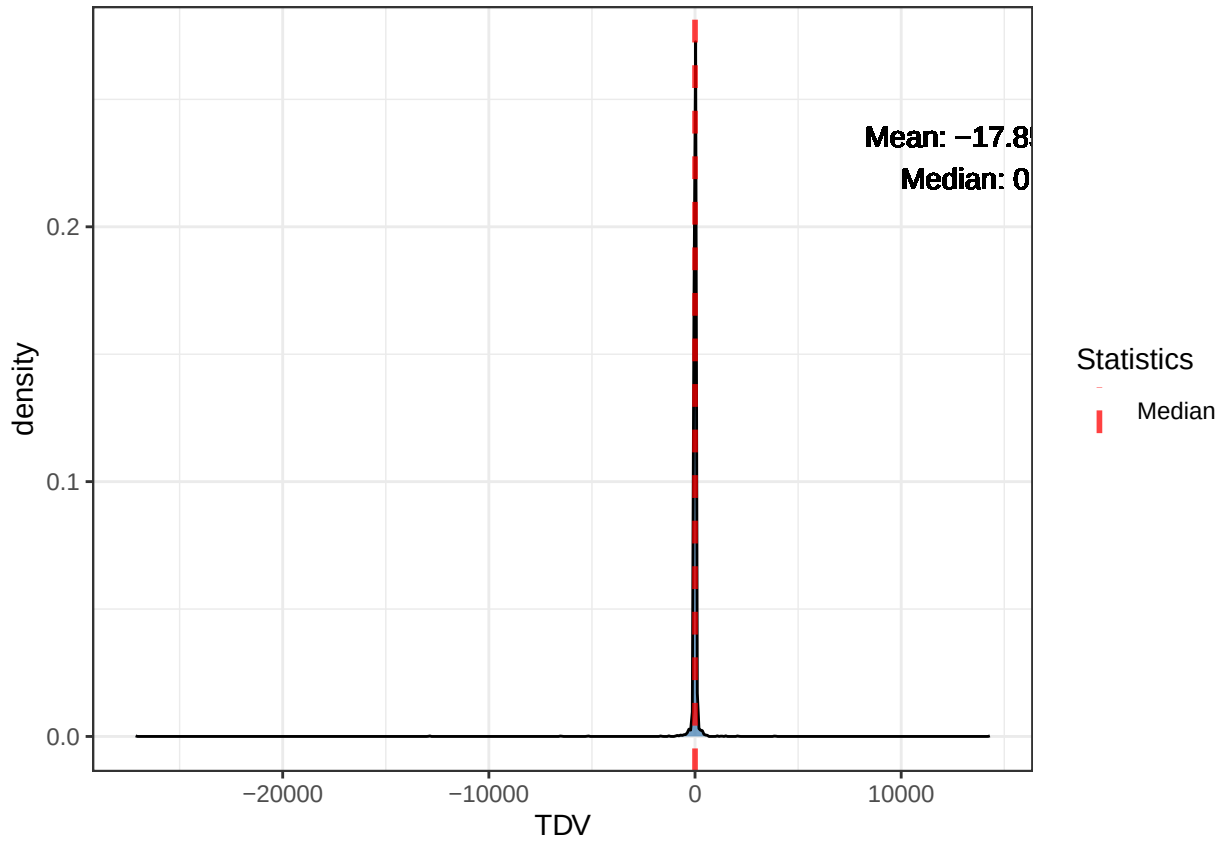
RQ1

TDD (Technical Debt Density) by PR Distribution



TDV (Technical Debt Variation)

TDV Distribution



Percentages for TDV < 0 (TD decrease), TDV = 0 (TD unchanged) and TDV > 0 (TD increased) by repo

```
## # A tibble: 12 x 4
##    repo              tdv_lower_than_zero tdv_equal_to_zero tdv_greater_than_zero
##    <chr>                           <dbl>             <dbl>                 <dbl>
##  1 accumulo                         26.2              47.0                  26.9
##  2 cayenne                          12.7              56.4                  30.9
##  3 commons-collecti~                35.7              53.6                  10.7
##  4 commons-io                       32.9              50.7                  16.4
##  5 commons-lang                     13.9              78.7                   7.38
##  6 helix                            19.2              37.0                  43.8
##  7 httpcomponents-c~                19.2              55.8                  25
##  8 maven-surefire                   24                68                     8
##  9 opennlp                          44.7              33.0                  22.3
## 10 struts                           37.1              36.4                  26.4
## 11 wicket                           10                57.5                  32.5
## 12 zookeeper                        17.6              52.9                  29.4
```

Mean of percentages of TDV (mean of repo percentages)

```
## # A tibble: 1 x 3
##   mean_tdv_lower_than_zero mean_tdv_equal_to_zero mean_tdv_greater_than_zero
##                      <dbl>                  <dbl>                      <dbl>
## 1                     24.4                   52.2                       23.3
```

Percentages of pre-existing TDV by repo

```
## # A tibble: 12 x 3
##    repo                  preexisting_tdv_equal_to_zero preexisting_tdv_greater~1
##    <chr>                                         <dbl>                     <dbl>
##  1 accumulo                                       57.6                      42.4
##  2 cayenne                                        75.9                      24.1
##  3 commons-collections                            53.8                      46.2
##  4 commons-io                                     62.5                      37.5
##  5 commons-lang                                   81.2                      18.8
##  6 helix                                          56.5                      43.5
##  7 httpcomponents-client                          69.5                      30.5
##  8 maven-surefire                                 68                        32
##  9 opennlp                                        44.0                      56.0
## 10 struts                                         51.8                      48.2
## 11 wicket                                         64.1                      35.9
## 12 zookeeper                                      67.6                      32.4
## # i abbreviated name: 1: preexisting_tdv_greater_than_zero
```

Mean of percentages of pre-existing TDV (mean of repo percentages)

```
## # A tibble: 1 x 2
##   preexisting_tdv_equal_to_zero mean_preexisting_tdv_greater_than_zero
##                           <dbl>                                  <dbl>
## 1                          62.7                                   37.3
```

Percentages for new TDV by repo

```
## # A tibble: 12 x 3
## # Groups:   repo [12]
##    repo                  new_tdv_equal_to_zero new_tdv_greater_than_zero
##    <chr>                                 <dbl>                     <dbl>
##  1 accumulo                               75.3                      24.7
##  2 cayenne                                90.9                       9.09
##  3 commons-collections                    66.7                      33.3
##  4 commons-io                             84                        16
##  5 commons-lang                           85                        15
##  6 helix                                  75.9                      24.1
##  7 httpcomponents-client                  94.7                       5.26
##  8 maven-surefire                        100                         0
##  9 opennlp                                95.8                       4.17
## 10 struts                                 91.4                       8.57
## 11 wicket                                 94.1                       5.88
## 12 zookeeper                              78.6                      21.4
```

Mean of percentages of new TDV (mean of repo percentages)

```
## # A tibble: 1 x 2
##   mean_new_tdv_equal_to_zero mean_new_tdv_greater_than_zero
##                        <dbl>                          <dbl>
## 1                       86.0                           14.0
```

RQ2

As data is unbalanced across repositories, as shown in the issue characterization (../issues_characterization.Rmd), we will examine the top 10 fixed issues and the top 10 unfixed issues for each repository, and then aggregate them into a rank using the *PositionScore* metric given by the following formula:

$$PositionScore_i = \sum_{j=1}^{k} Position_j$$

given a rule i, its $PositionScore_i$ will be the sum of its position across all k repositiries.After calculating the metrics for all rules, we rank them in a TOP 10 of the most frequent rules (fixed or unfixed, depending on the context) across repositories. The lower the PositionScore, the more frequent the rule across repositories.

TOP 10 fixed per repo ranked by count

All top 10 fixed rules for each repo.

```
## # A tibble: 10 x 13
## # Groups:   rank [10]
##     rank accumulo      cayenne `commons-collections` `commons-io` `commons-lang`
##    <int> <chr>         <chr>   <chr>                 <chr>        <chr>
##  1     1 java:S117  -~ java:S~ java:S112  -  10      java:S4719 ~ java:S4165  -~
##  2     2 java:S112  -~ java:S~ java:S1125  -  8      java:S1192 ~ java:S1192  -~
##  3     3 java:S1192  ~ java:S~ java:S1170  -  6      java:S4042 ~ java:S125  - ~
##  4     4 java:S2589  ~ java:S~ java:S3008  -  5      java:S899  ~ java:S2293  -~
##  5     5 java:S1125  ~ java:S~ java:S1874  -  5      java:S1125 ~ java:S3878  -~
##  6     6 java:S101  -~ java:S~ java:S1197  -  5      java:S2129 ~ java:S1612  -~
##  7     7 java:S3776  ~ java:S~ java:S1192  -  4      java:S1874 ~ java:S100  - ~
##  8     8 java:S2293  ~ java:S~ java:S1134  -  4      java:S2293 ~ java:S1611  -~
##  9     9 java:S1155  ~ java:S~ java:S125  -  4       java:S1612 ~ java:S4968  -~
## 10    10 java:S1135  ~ java:S~ java:S1126  -  2      java:S1481 ~ java:S3358  -~
## # i 7 more variables: helix <chr>, `httpcomponents-client` <chr>,
## #   `maven-surefire` <chr>, opennlp <chr>, struts <chr>, wicket <chr>,
## #   zookeeper <chr>
```

TOP 10 fixed

TOP 10 fixed rules by PositionScore metric.

```
## # A tibble: 57 x 2
##    rule       position_score
##    <chr>               <dbl>
##  1 java:S1192             41
##  2 java:S1874             84
##  3 java:S112              90
```

```
##  4 java:S2293             94
##  5 java:S3776            102
##  6 java:S100             103
##  7 java:S1161            107
##  8 java:S3740            108
##  9 java:S1125            111
## 10 java:S117             113
## # i 47 more rows
```

TOP 10 unfixed per repo ranked by count

All top 10 unfixed rules for each repo.

```
## # A tibble: 10 x 13
## # Groups:   rank [10]
##     rank accumulo      cayenne ‘commons-collections‘ ‘commons-io‘ ‘commons-lang‘
##    <int> <chr>         <chr>   <chr>                 <chr>        <chr>
##  1     1 java:S117  -~ java:S~ java:S1192  -  336    java:S1192 ~ java:S1192  -~
##  2     2 java:S2589  ~ java:S~ java:S125  -  215     java:S100  ~ java:S100  - ~
##  3     3 java:S1192  ~ java:S~ java:S108  -  174     java:S899  ~ java:S1168  -~
##  4     4 java:S101  -~ java:S~ java:S1117  -  122    java:S4042 ~ java:S1133  -~
##  5     5 java:S112  -~ java:S~ java:S2293  -  93     java:S4719 ~ java:S3776  -~
##  6     6 java:S1125  ~ java:S~ java:S1125  -  89     java:S2184 ~ java:S4165  -~
##  7     7 java:S3776  ~ java:S~ java:S2160  -  57     java:S117  ~ java:S125  - ~
##  8     8 java:S2293  ~ java:S~ java:S128  -  52      java:S1133 ~ java:S1172  -~
##  9     9 java:S116  -~ java:S~ java:S3776  -  52     java:S108  ~ java:S1149  -~
## 10    10 java:S2142  ~ java:S~ java:S2065  -  51     java:S3878 ~ java:S3878  -~
## # i 7 more variables: helix <chr>, ‘httpcomponents-client‘ <chr>,
## #   ‘maven-surefire‘ <chr>, opennlp <chr>, struts <chr>, wicket <chr>,
## #   zookeeper <chr>
```

TOP 10 unfixed

TOP 10 unfixed rules by PositionScore metric.

```
## # A tibble: 50 x 2
##    rule       position_score
##    <chr>               <dbl>
##  1 java:S1192             22
##  2 java:S3776             85
##  3 java:S100             100
##  4 java:S106             100
##  5 java:S117             101
##  6 java:S1133            103
##  7 java:S1874            107
##  8 java:S3740            108
##  9 java:S1135            111
## 10 java:S1125            112
## # i 40 more rows
```

Outliers

Higher TDD

```
## # A tibble: 1,959 x 5
##    repo     pr_number total_debt total_ncloc   tdd
##    <chr>        <int>      <int>       <int> <dbl>
##  1 wicket         676        600          58 10.3
##  2 helix         1975        261          33  7.91
##  3 wicket         708       3157         432  7.31
##  4 accumulo      2922        410          69  5.94
##  5 accumulo      4215        140          25  5.6
##  6 wicket         709       2265         474  4.78
##  7 opennlp        563       1035         232  4.46
##  8 opennlp        561      74742       17056  4.38
##  9 wicket         351       2159         599  3.60
## 10 opennlp        569       1161         325  3.57
## # i 1,949 more rows
```

Higher TDV

```
## # A tibble: 1,959 x 3
##    pr_number repo                  tdv
##        <int> <chr>                <int>
##  1      2022 accumulo             14300
##  2      1946 accumulo              3875
##  3      2503 accumulo              2078
##  4      2335 accumulo              1492
##  5      1891 accumulo              1286
##  6      1605 accumulo              1082
##  7      4054 accumulo               593
##  8       427 struts                 579
##  9       483 httpcomponents-client  525
## 10       559 struts                 510
## # i 1,949 more rows
```

Lower TDV

```
## # A tibble: 1,959 x 3
##    pr_number repo                   tdv
##        <int> <chr>                <int>
##  1       799 struts               -27133
##  2       561 opennlp              -12880
##  3      3604 accumulo              -6508
##  4      3402 accumulo              -5186
##  5       670 struts                -1655
##  6      3117 accumulo              -1262
##  7      3229 accumulo               -899
##  8       484 httpcomponents-client  -884
##  9       431 opennlp                -813
## 10      2705 accumulo               -724
## # i 1,949 more rows
```

Extra

Number of PRs with pre-existing TD

```
##      n
## 1 1931
```

Number of java:S1192 issues that affect test code

```
##       n
## 1 18504
```

Number of java:S1192 issues that affect not test code

```
##      n
## 1 8346
```

Summary NCLOC

```
##      ncloc
##  Min.   :     3.0
##  1st Qu.:   256.5
##  Median :   651.0
##  Mean   :  2327.5
##  3rd Qu.:  1772.5
##  Max.   :240475.0
```