Tianyu Qiu
767920912
INFOSYS 750
7 November 2019

GitHub Open Source Software Development: Influencing Factors and Correlation Analysis

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

Abstract: *The contribution of this report is to explore the main factors that influence open-source software development and find the best model to understand evolution in some of the OSS projects. A total of 4 critical factors affecting open-source software development were found in this report and found the best model.*

Table of Contents

I Introduction

GitHub is a web-based Git that provides version-control storage and hosting (Dabbish et al., 2012). According to the annual report released by GitHub in 2015, the GitHub community has reached 24 million developers, 1.5 million organizations including global technology giants such as Microsoft, Facebook, Google and Apple, and more than 67 million resource database information (Blischak et al., 2016). Since September 2016, the number of submissions has reached 1 billion, and the number is proliferation (Blischak et al., 2016). Projects hosted on GitHub can be accessed and operated with standard Git commands, and GitHub offers a range of social networking features, such as following and commenting. Up to now, GitHub has more than 12 million open source projects, and the number is still growing. The analysis of influencing factors and correlation in the development process can reveal the development level of GitHub open source software and the progress of the project to some extent (Kalliamvakou et al., 2015). By analyzing the data generated by open-source software during development, we proposed eight significant factors affecting development quality: Time, Forks, Members, Commits, Issues, Watchers, PullReq, CommitCmnt, and analyzed the correlation among these factors. Section 1 gives a brief introduction to GitHub, focusing on the factors that influence the development process of GitHub open-source software. Section 2 proposes the Research Question. Section 3 defines the main variables and explores the data visually. Section 4 completes the cleaning and preparation of the data. Section 5 constructed the Longitudinal Multi-level Model and analyzed the results in detail. Section 6 explains the superior results and offers a few suggestions for developers. Finally, the paper summarizes and looks forward to the follow-up work.

II Research Questions

Many researchers have paid attention to the influential factors in the development process of open source software services (Perez-Riverol et al., 2016), but the correlation analysis of these influential factors has not received enough attention. Correlation analysis of influencing factors can help contributors of open-source software better participate in the development and maintenance of software, and also improve the efficiency and quality of open-source software development (Tsay et al., 2014). For example, whether it is possible to accelerate the problem-solving speed in the development process of open source software by adjusting one or several

influential factors in the development process of open-source software. Such questions have become the primary motivation of this paper. What influences can we adjust to make the development of open-source software faster and better? How to adjust? Given this, this paper proposes the following research questions: (1) Whether there is a correlation between the significant factors that affect the development process of GitHub open-source software, and if so, what is the correlation? (2) How to make use of the correlation of influential factors in the development process of GitHub open-source software to guide the development of open-source software better?

III Definition of Main Variables & Visual Exploration

3.1 Definition of Main Variables

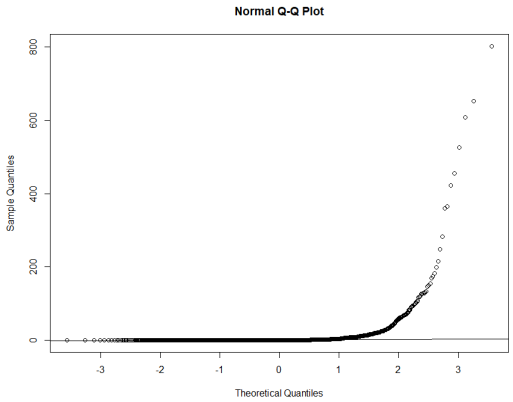
Table 1 Definition of Main Variables

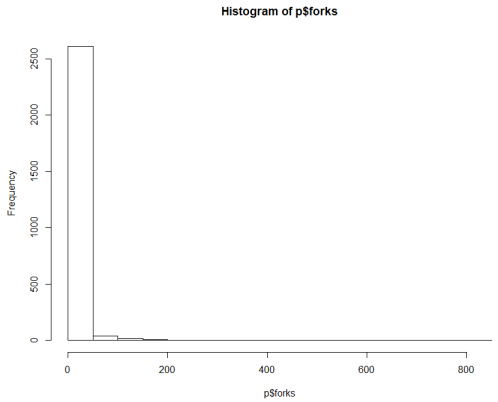
Variable	Definition
PrjID	A unique id number for each project
Period	Represents the current record contains data for which period of year
Time	A sequence for time of observations
SatrtDate	Beginning of observation
EndDate	End of observation
Forks	Number of times a project is Forksed
Members	Number of members
Commits	Number of coding activities
Issues	Number of problem/bugs raised or requests for new features
Watchers	Number of people interested in project
PullReq	Number of code changes request for review.
CommitCmnt	Number of discussion on commits

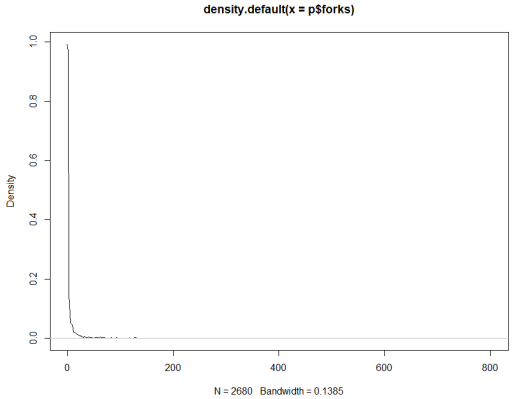
Since there are inclusion and inclusion relationships in all 31 variables, such as PullReqCmnt is a subset of CommitCmnt, MemIssue is a subset of Issues, so we only focus on the top 8 variables in this report: Time, Forks, Members, Commits, Issues, Watchers, PullReq and CommitCmnt.

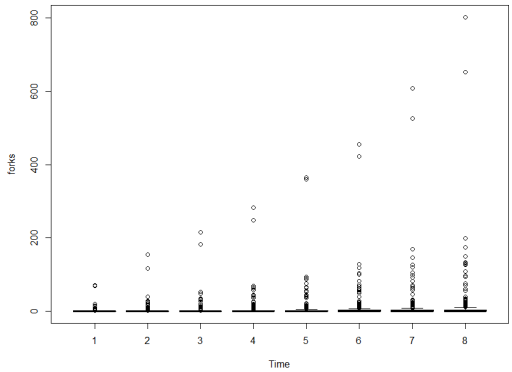
```
> describe(p)
          vars    n       mean         sd   median    trimmed       mad  min      max    range  skew kurtosis       se
prjId        1 2680 2870787.14 4395980.89 708445.0 2122170.43 955768.47 2647 13095415 13092768  1.38     0.03 84915.78
Time         2 2680       4.50       2.29      4.5       4.50      2.97    1        8        7  0.00    -1.24     0.04
forks        3 2680       5.92      34.20      0.0       0.81      0.00    0      802      802 13.92   243.10     0.66
members      4 2680      11.75      12.54      8.0       9.16      4.45    0       97       97  2.93     9.85     0.24
commits      5 2680     155.67     538.10     16.0      53.81     23.72    0    10706    10706  9.14   117.57    10.39
issues       6 2680      24.85     112.94      0.0       2.20      0.00    0     2139     2139  8.50    98.21     2.18
watchers     7 2680      67.09     331.49      2.0       7.51      2.97    0     5153     5153  8.57    90.87     6.40
pullReq      8 2680      85.01    1050.15      0.0       1.68      0.00    0    28626    28626 21.73   522.59    20.29
CmtCmnt      9 2680       2.77      32.72      0.0       0.00      0.00    0      788      788 19.21   403.71     0.63
```

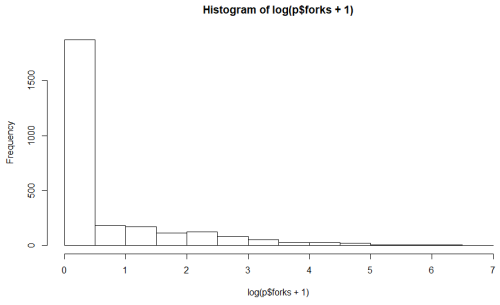
3.2 Visual Exploration





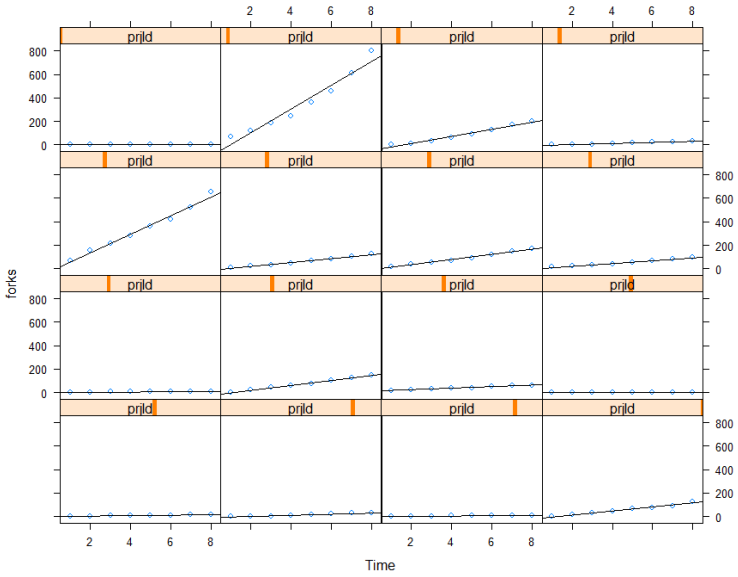




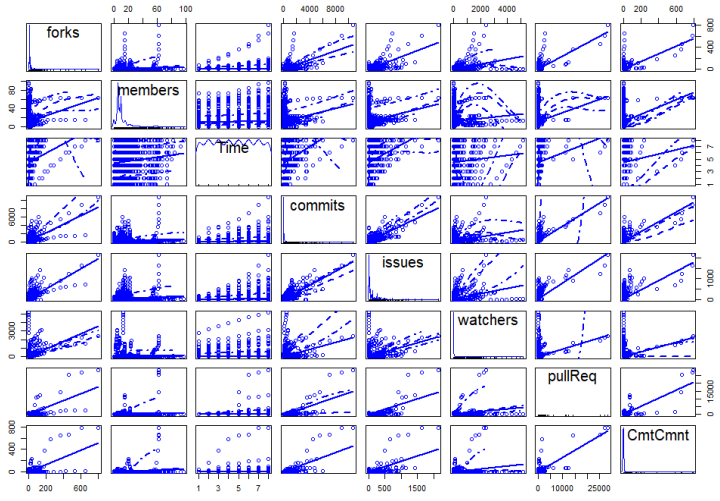



```
> lillie.test(p$forks)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  p$forks
D = 0.43129, p-value < 2.2e-16
```

First, we use the graph method to test the normal distribution of the dependent variable Forks. By observing the graph, we guess that Forks are not normally distributed. Subsequently, KS Test was carried out on Forks using statistics. The calculated results showed that, $p < 0.05$, reject H0, so Forks did not meet the normal distribution.



The purpose of this paper is to explore the influence of different factors on Forks in the time dimension, so we selected 16 projects (5% of the total number of projects) to observe the relationship between Forks and Time, and found that Forks and Time are linearly related.

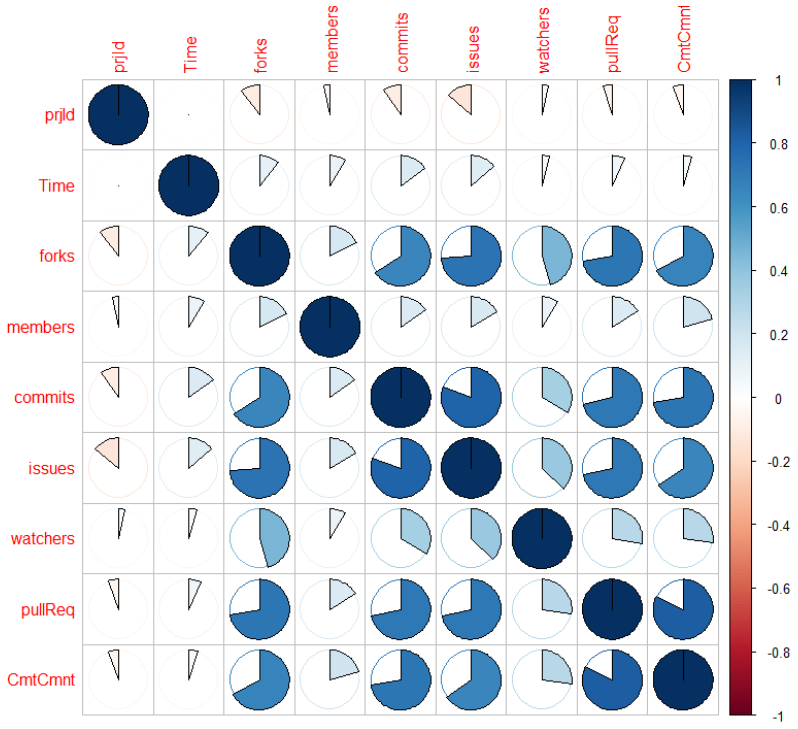


We have conducted Linearity Check for eight main variables. The result shows that there is a linear relationship between 8 main variables. The Linearity assumption is satisfied.

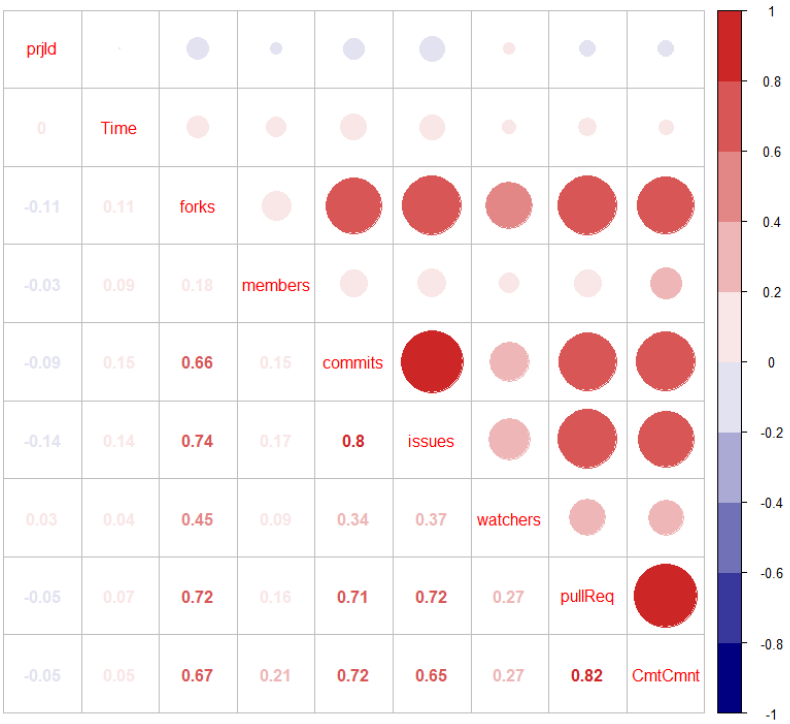
IV Data Cleaning & Preparation

4.1 Importance Analysis

First, we need to find the covariates suitable for study. The method is to find out the connection between the primary variable except for the dependent variable (Forks) and the dependent variable (Forks) and sort them according to the importance degree of correlation to complete the preliminary screening of covariables.



We performed a visual analysis of the correlation between the eight significant variables. It was observed that variables significantly correlated with Forks were Commits, Issues, PullReq, and CommitCmnt. However, graphical methods cannot accurately describe the degree of connection, so we use more accurate methods to sort.



As shown in the figure above, the correlation between significant variables and Forks can be sorted in order of importance: Issues (0.74), PullReq (0.72), CommitCmnt (0.67), Commits (0.66), Watchers (0.45), Members (0.18).

Watchers and member were discarded because the correlation coefficient was less than 0.5. Finally, Issues, PullReq, CommitCmnt, and Commits were selected as covariables for the study.

4.2 Data Cleaning

```
> sum(is.null(p)) > sum(is.na(p))
[1] 0             [1] 0
```

Using is. null () function and is.na() function, we find that there is no null value and missing value in the main variables.

V Multi-level Longitudinal Model

5.1 Model A

```
> summary(model.a)
Linear mixed-effects model fit by REML
 Data: p
       AIC      BIC    logLik
  24518.55 24536.23 -12256.28

Random effects:
 Formula: ~1 | prjId
        (Intercept) Residual
StdDev:    28.05296 19.62183

Fixed effects: forks ~ 1
               Value Std.Error   DF  t-value p-value
(Intercept) 5.920522  1.578868 2345 3.749854   2e-04

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-13.58635756  -0.01738895  -0.01738895  -0.01738895  23.77000317

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.a)
prjId = pdLogChol(1)
            Variance StdDev
(Intercept) 786.9686 28.05296
Residual    385.0160 19.62183
```

Composite Model A : Forks = 5.92 + e

Estimate of fixed effects: the initial status of Forks at the Time 0 is 5.92(1.58) at 0.05 level of significance. (p-value = 0 < 0.05)

Variance components:

Level 1 (within project variance) gets the estimate of 385.02(19.62)

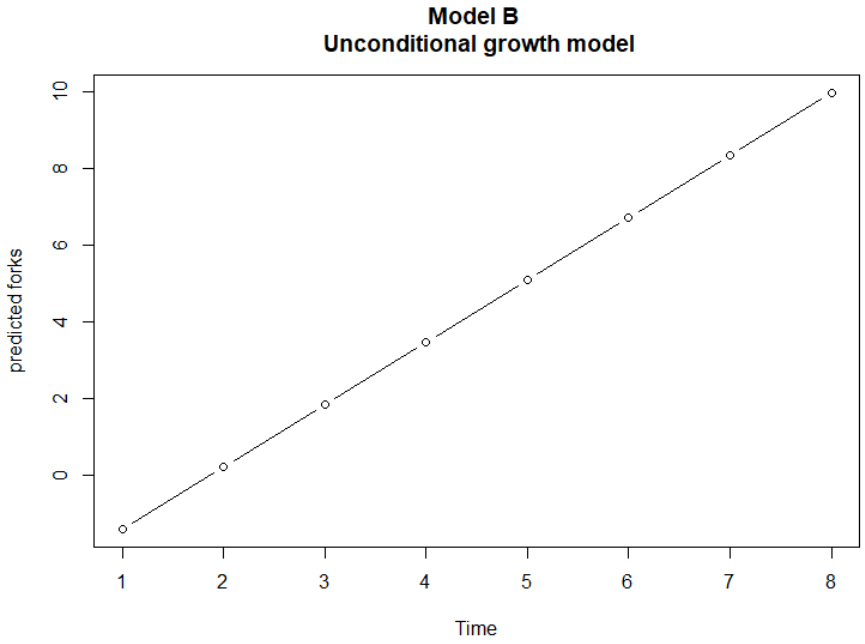
Level 2 (between project variance) receives the estimate of 786.87(28.05)

ICC = 786.87 / (385.02 + 786.87) = 0.6715

Therefore, 67.15% variation in the Forks is attribute to differences among projects.

5.2 Model B

```
> summary(model.b)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  17339.53 17374.89 -8663.765

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 6.379360 (Intr)
Time        7.676154 -0.915
Residual    3.954964

Fixed effects: forks ~ Time
               Value Std.Error   DF   t-value p-value
(Intercept) -1.390299 0.3872231 2344 -3.590433   3e-04
Time         1.624627 0.4208734 2344  3.860132   1e-04
 Correlation:
     (Intr)
Time -0.852

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-15.523533370  -0.015230553   0.001421493   0.023624222  24.760982473

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.b)
prjId = pdLogChol(Time)
            Variance StdDev   Corr
(Intercept) 40.69623 6.379360 (Intr)
Time        58.92333 7.676154 -0.915
Residual    15.64174 3.954964
```



Composite Model B:

Level1: Forks = a + b* Time + j

Level 2: a = -1.39 + y_0i

b = 1.62 + y_1i

Forks = -1.39 + 1.62 * Time + e , where e = y_0i + y_1i * Time + j

Estimates of fixed effects show significant values. From the result, we can see that the rate of change is the estimates of Time. The estimated rate of change in Forks for projects is -1.39 (p-value < 0.05) where the estimated initial test score is 1.62 (p-value = 0<0.05). Therefore, we can interpret this estimate: there is an increase in Forks over time from one period to the next, there will be an increase of 1.62 in Forks each period.

5.3 Model C

5.3.1 Model C1

```
> summary(model.c1)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  16068.64 16115.79 -8026.32

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 4.574520 (Intr)
Time        6.114526 -0.65
Residual    2.946932

Fixed effects: forks ~ commits * Time
                 Value Std.Error   DF  t-value p-value
(Intercept)   0.3982287 0.2835573 2342  1.40440  0.1603
commits      -0.0035673 0.0009048 2342 -3.94265  0.0001
Time          0.6995141 0.3364829 2342  2.07890  0.0377
commits:Time  0.0033042 0.0000870 2342 37.98042  0.0000
 Correlation:
              (Intr) commts Time
commits       -0.078
Time          -0.605 -0.051
commits:Time  0.161 -0.543 -0.031

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-11.150774466  -0.054802341  -0.008418021   0.020645215  17.507008623

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.c1)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 20.926232 4.574520 (Intr)
Time        37.387423 6.114526 -0.65
Residual     8.684405 2.946932
```

The fixed effects values are presenting the significant impact of Commits on Forks both at initial status(estimate = -0.003 at 0.05 l.o.s) and over time (estimate = 0.003 at 0.05 l.o.s).The estimate initial Forks for projects with Commits is 0.398 at 0.05 level of significance.

The estimate at the rate of change in Forks for projects with Commits is 0.699. Then, there is no significant gap between Forks at the initial status and in the rate of change of Commits.

Rseudo R2 is (40.696-20.926)/40.696 = 0.4858, which means approx. 48.58% of between projects variance in Forks is associated with Commits*Time.

5.3.2 Model C2

```
> summary(model.c2)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15826.99 15874.14 -7905.494

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 4.764826 (Intr)
Time        5.653387 -0.613
Residual    2.790087

Fixed effects: forks ~ issues * Time
                Value  Std.Error   DF   t-value p-value
(Intercept) 0.1150230 0.28811547 2342  0.399225  0.6898
issues      0.0228351 0.00529724 2342  4.310761  0.0000
Time        0.7375672 0.31103264 2342  2.371350  0.0178
issues:Time 0.0130036 0.00042976 2342 30.258143  0.0000
 Correlation:
            (Intr) issues Time
issues      -0.006
Time        -0.585 -0.060
issues:Time  0.081 -0.715  0.005

Standardized Within-Group Residuals:
         Min            Q1           Med            Q3           Max
-10.948531873  -0.030742924  -0.006935286   0.010920443  15.268460513

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.c2)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 22.703569 4.764826 (Intr)
Time        31.960785 5.653387 -0.613
Residual     7.784585 2.790087
```

The fixed effects values are presenting the significant impact of Issues on Forks both at initial status(estimate = 0.022 at 0.05 l.o.s) and over time (estimate = 0.013 at 0.05 l.o.s). The initial estimate Forks for projects with Issues is 0.115 at 0.05 level of significance.

The estimate at the rate of change in Forks for projects with Issues is 0.738. Then, there is no significant gap between Forks at the initial status and in the rate of change of Issues.

Rseudo R2 is (40.696-22.703)/40.696 = 0.4421, which means approx. 44.21% of between projects variance in Forks is associated with Issues *Time.

5.3.3 Model C3

```
> summary(model.c3)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15347.79 15394.94 -7665.895

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 4.176530 (Intr)
Time        6.370656 -0.53
Residual    2.455937

Fixed effects: forks ~ pullReq * Time
                 Value Std.Error   DF   t-value p-value
(Intercept)  -0.7790437 0.2514587 2342  -3.09810   0.002
pullReq      -0.0130439 0.0003633 2342 -35.90262   0.000
Time          1.4184282 0.3489843 2342   4.06445   0.000
pullReq:Time  0.0026100 0.0000485 2342  53.81202   0.000
 Correlation:
             (Intr) pullRq Time
pullReq      -0.007
Time         -0.503 -0.004
pullReq:Time  0.028 -0.885 -0.004

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-11.904974458  -0.016208954   0.008217791   0.026537850  14.006642711

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.c3)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 17.443401 4.176530 (Intr)
Time        40.585255 6.370656 -0.53
Residual     6.031628 2.455937
```

The fixed effects values are presenting the significant impact of PullReq on Forks both at initial status(estimate = -0.013 at 0.05 l.o.s) and over time (estimate = 0.002 at 0.05 l.o.s). The estimate initial Forks for projects with Issues is -0.779 at 0.05 level of significance.

The estimate at the rate of change in Forks for projects with PullReq is 1.418. Then, there is no significant gap between Forks at the initial status and in the rate of change of PullReq.

Rseudo R2 is (40.696-17.443)/40.696 = 0.5713, which means approx. 57.13% of between projects variance in Forks is associated with PullReq *Time.

5.3.4 Model C4

```
> summary(model.c4)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15710.01 15757.16 -7847.007

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 3.977967 (Intr)
Time        6.929991 -0.544
Residual    2.661315

Fixed effects: forks ~ CmtCmnt * Time
                  Value Std.Error   DF   t-value p-value
(Intercept)  -0.8693983 0.2460883 2342  -3.53287   4e-04
CmtCmnt      -0.3958271 0.0104929 2342 -37.72344   0e+00
Time          1.5148047 0.3798429 2342   3.98798   1e-04
CmtCmnt:Time  0.0668253 0.0020711 2342  32.26602   0e+00
 Correlation:
             (Intr) CmtCmn Time
CmtCmnt       0.020
Time         -0.506 -0.022
CmtCmnt:Time  0.080  0.106 -0.033

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-13.83445327  -0.02179021   0.01231624   0.03842108  15.10875829

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.c4)
prjId = pdLogChol(Time)
            Variance StdDev   Corr
(Intercept) 15.82422 3.977967 (Intr)
Time        48.02477 6.929991 -0.544
Residual     7.08260 2.661315
```

The fixed effects values are presenting the significant impact of CommitCmnt on Forks both at initial status(estimate = -0.395 at 0.05 l.o.s) and over time (estimate = 0.066 at 0.05 l.o.s). The estimate initial Forks for projects with Issues is -0.869 at 0.05 level of significance.

The estimate at the rate of change in Forks for projects with CommitCmnt is 1.514. Then, there is no significant gap between Forks at the initial status and in the rate of change of CommitCmnt.

Rseudo R2 is (40.696-15.824)/40.696 = 0.6111, which means approx. 61.11% of between projects variance in Forks is associated with CommitCmnt *Time.

5.4 Model D

5.4.1 Model D1

```
> summary(model.d1)
Linear mixed-effects model fit by maximum likelihood
 Data: p
      AIC      BIC    logLik
  15752.3 15811.24 -7866.152

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 4.636934 (Intr)
Time        5.581037 -0.613
Residual    2.752705

Fixed effects: forks ~ commits * Time + issues * Time
                  Value  Std.Error   DF   t-value p-value
(Intercept)   0.3133315 0.28306017 2340  1.106943  0.2684
commits      -0.0003662 0.00096723 2340 -0.378629  0.7050
Time          0.6019155 0.30773653 2340  1.955944  0.0506
issues        0.0253397 0.00552577 2340  4.585735  0.0000
commits:Time  0.0011944 0.00015730 2340  7.593078  0.0000
Time:issues   0.0085848 0.00070329 2340 12.206531  0.0000
 Correlation:
             (Intr) commts Time   issues cmmt:T
commits      -0.086
Time         -0.579 -0.033
issues        0.026 -0.326 -0.048
commits:Time  0.111 -0.562 -0.024  0.217
Time:issues  -0.039  0.411  0.025 -0.569 -0.797

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-10.973733451  -0.041956047  -0.008480027   0.015014923  15.091520507

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.d1)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 21.501160 4.636934 (Intr)
Time        31.147977 5.581037 -0.613
Residual     7.577386 2.752705
```

The p-value of Commits is 0.705 (0.705 > 0.05 at 0.05 l.o.s), thus Commits have no significant effect on Forks；The p-value of Commits is 0.0 (0.705 > 0.05 at 0.05 l.o.s), thus Commits have no significant effect on Forks；Issues have a positive impact on Forks. Over time, both Issues and Commits have a positive impact on the Forks. Therefore, the impact of Commits on Forks depends on Time.

5.4.2 Model D2

```
> summary(model.d2)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15221.89 15280.83 -7600.946

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 4.060848 (Intr)
Time        6.014549 -0.507
Residual    2.405931

Fixed effects: forks ~ commits * Time + pullReq * Time
                  Value Std.Error   DF    t-value p-value
(Intercept)  -0.4034461 0.2491340 2340  -1.619394  0.1055
commits       0.0017254 0.0008261 2340   2.088693  0.0368
Time          1.1144335 0.3307555 2340   3.369357  0.0008
pullReq      -0.0107172 0.0004211 2340 -25.448009  0.0000
commits:Time  0.0008876 0.0001087 2340   8.165758  0.0000
Time:pullReq  0.0021320 0.0000675 2340  31.578667  0.0000
 Correlation:
             (Intr) commts Time   pullRq cmmt:T
commits      -0.104
Time         -0.480 -0.034
pullReq       0.087 -0.152 -0.036
commits:Time  0.183 -0.569 -0.042  0.508
Time:pullReq -0.106  0.235  0.039 -0.906 -0.685

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-12.282433362  -0.010005208   0.001547615   0.012530151  14.618388752

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.d2)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 16.490487 4.060848 (Intr)
Time        36.174801 6.014549 -0.507
Residual     5.788503 2.405931
```

It is worth noting that PullReq has a negative impact on the Forks, but over time, PullReq will have a positive impact on the Forks.

5.4.3 Model D3

```
> summary(model.d3)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15468.74 15527.67 -7724.368

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 3.895711 (Intr)
Time        6.512353 -0.531
Residual    2.538589

Fixed effects: forks ~ commits * Time + CmtCmnt * Time
                 Value Std.Error   DF    t-value p-value
(Intercept)  -0.3089319 0.2440123 2340  -1.266051  0.2056
commits       0.0009558 0.0008448 2340   1.131489  0.2580
Time          1.1112029 0.3581008 2340   3.103045  0.0019
CmtCmnt      -0.3015236 0.0120943 2340 -24.931067  0.0000
commits:Time  0.0013800 0.0001093 2340  12.624443  0.0000
Time:CmtCmnt  0.0432264 0.0025575 2340  16.901750  0.0000
 Correlation:
              (Intr) commts Time   CmtCmn cmmt:T
commits       -0.110
Time          -0.495 -0.030
CmtCmnt        0.119 -0.252 -0.045
commits:Time   0.186 -0.559 -0.039  0.556
Time:CmtCmnt -0.048  0.165  0.010 -0.277 -0.592

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-12.844443583  -0.010815647  -0.001123144   0.014684706  15.560007013

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.d3)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 15.176563 3.895711 (Intr)
Time        42.410743 6.512353 -0.531
Residual     6.444435 2.538589
```

Commits had no significant effect on the Forks ($p > 0.05$ l.o.s). However, over time, Commits have a weak positive impact on the Forks. CommitCmnt has a negative impact on Forks, but over time, CommitCmnt has a positive impact on Forks.

5.4.4 Model D4

```
> summary(model.d4)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15151.71 15210.65 -7565.856

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 4.068207 (Intr)
Time        5.727441 -0.504
Residual    2.383497

Fixed effects: forks ~ issues * Time + pullReq * Time
                 Value  Std.Error   DF    t-value p-value
(Intercept)  -0.4054496 0.24687905 2340  -1.642301  0.1007
issues        0.0320967 0.00466989 2340   6.873127  0.0000
Time          1.0714939 0.31488579 2340   3.402802  0.0007
pullReq      -0.0099131 0.00044679 2340 -22.187075  0.0000
issues:Time   0.0031524 0.00052350 2340   6.021758  0.0000
Time:pullReq  0.0019852 0.00007253 2340  27.371059  0.0000
 Correlation:
             (Intr) issues Time   pullRq isss:T
issues       -0.017
Time         -0.483 -0.047
pullReq       0.072 -0.114 -0.034
issues:Time   0.110 -0.607 -0.020  0.549
Time:pullReq -0.079  0.136  0.036 -0.920 -0.677

Standardized Within-Group Residuals:
         Min            Q1           Med            Q3           Max
-12.439308652  -0.006571883   0.001011118   0.006698369  15.081212748

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.d4)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 16.550311 4.068207 (Intr)
Time        32.803577 5.727441 -0.504
Residual     5.681056 2.383497
```

Issues have a positive impact on Forks, and PullReq has a negative impact on Forks. Over time, the positive impact of Issues on Forks will diminish, and PullReq will have a positive impact on Forks.

5.4.5 Model D5

```
> summary(model.d5)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15307.19 15366.13 -7643.597

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 3.920849 (Intr)
Time        6.094398 -0.463
Residual    2.452174

Fixed effects: forks ~ issues * Time + CmtCmnt * Time
                 Value Std.Error   DF    t-value p-value
(Intercept)  -0.2786892 0.2409778 2340  -1.156493  0.2476
issues        0.0196382 0.0047607 2340   4.125072  0.0000
Time          1.0750729 0.3350592 2340   3.208606  0.0014
CmtCmnt      -0.2654416 0.0116743 2340 -22.737288  0.0000
issues:Time   0.0065903 0.0004771 2340  13.813819  0.0000
Time:CmtCmnt  0.0397460 0.0024214 2340  16.414595  0.0000
 Correlation:
              (Intr) issues Time   CmtCmn isss:T
issues       -0.015
Time         -0.442 -0.046
CmtCmnt       0.081 -0.088 -0.044
issues:Time   0.101 -0.613 -0.013  0.491
Time:CmtCmnt -0.006 -0.003  0.003 -0.263 -0.464

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-1.235430e+01 -3.787099e-03 -8.024992e-04  1.435950e-03  1.570997e+01

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.d5)
prjId = pdLogChol(Time)
            Variance StdDev   Corr
(Intercept) 15.37306 3.920849 (Intr)
Time        37.14168 6.094398 -0.463
Residual     6.01316 2.452174
```

Issues have a positive impact on Forks, and CommitCmnt has a negative impact on Forks. Over time, the positive impact of Issues on Forks will diminish, and CommitCmnt will have a positive impact on Forks.

5.4.6 Model D6

```
> summary(model.d6)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15174.78 15233.72 -7577.391

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 4.018639 (Intr)
Time        6.333572 -0.455
Residual    2.351732

Fixed effects: forks ~ pullReq * Time + CmtCmnt * Time
                  Value Std.Error   DF    t-value p-value
(Intercept)  -0.7179701 0.2425423 2340  -2.960185  0.0031
pullReq      -0.0104386 0.0004228 2340 -24.688781  0.0000
Time          1.4105641 0.3473896 2340   4.060467  0.0001
CmtCmnt      -0.1566389 0.0186856 2340  -8.382863  0.0000
pullReq:Time  0.0018205 0.0000860 2340  21.169725  0.0000
Time:CmtCmnt  0.0385815 0.0028200 2340  13.681316  0.0000
 Correlation:
              (Intr) pullRq Time   CmtCmn pllR:T
pullReq       -0.023
Time          -0.437  0.012
CmtCmnt        0.048 -0.545 -0.035
pullReq:Time   0.040 -0.871 -0.023  0.806
Time:CmtCmnt   0.018  0.454 -0.002 -0.601 -0.677

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-12.489320608  -0.016590689   0.008549022   0.027403805  15.327856160

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.d6)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 16.149458 4.018639 (Intr)
Time        40.114132 6.333572 -0.455
Residual     5.530644 2.351732
```

PullReq and CommitCmnt have a negative impact on Forks. Over time, both will have a positive impact on the Forks.

5.5 Model E

In Model E, we selected two variables, Issues(0.74) and PullReq(0.72), which are most relevant to Forks for analysis.

5.5.1 Model E1

```
> summary(model.e1)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15184.19 15237.23 -7583.096

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 4.081529 (Intr)
Time        5.744571 -0.555
Residual    2.413289

Fixed effects: forks ~ issues + pullReq * Time
                  Value  Std.Error   DF   t-value p-value
(Intercept)  -0.5709514 0.24652570 2341  -2.31599  0.0206
issues        0.0498359 0.00374130 2341  13.32045  0.0000
pullReq      -0.0113354 0.00037783 2341 -30.00131  0.0000
Time          1.1066714 0.31572843 2341   3.50514  0.0005
pullReq:Time  0.0022646 0.00005390 2341  42.01912  0.0000
 Correlation:
             (Intr) issues pullRq Time
issues        0.064
pullReq       0.014  0.328
Time         -0.530 -0.074 -0.028
pullReq:Time -0.005 -0.467 -0.893  0.031

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-12.458391403  -0.010834014   0.004275173   0.015607062  15.414196106

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.e1)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 16.658877 4.081529 (Intr)
Time        33.000101 5.744571 -0.555
Residual     5.823962 2.413289
```

Issues have a negative impact on the Forks. In the beginning, PullReq has a negative impact on the Forks, but over time, PullReq will have a positive impact on the Forks.

5.5.2 Model E2

```
> summary(model.e2)
Linear mixed-effects model fit by maximum likelihood
 Data: p
       AIC      BIC    logLik
  15785.21 15838.26 -7883.607

Random effects:
 Formula: ~Time | prjId
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 4.578487 (Intr)
Time        5.481954 -0.567
Residual    2.773033

Fixed effects: forks ~ issues * Time + pullReq
                 Value  Std.Error   DF   t-value p-value
(Intercept) 0.1194283 0.27864377 2341  0.428606  0.6682
issues      0.0224435 0.00526908 2341  4.259469  0.0000
Time        0.7343130 0.30174115 2341  2.433586  0.0150
pullReq     0.0013656 0.00020259 2341  6.740572  0.0000
issues:Time 0.0123523 0.00044337 2341 27.859846  0.0000
 Correlation:
            (Intr) issues Time   pullRq
issues      -0.007
Time        -0.541 -0.062
pullReq     -0.001  0.028 -0.003
issues:Time  0.080 -0.701  0.006 -0.262

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-12.559417209  -0.028393274  -0.006555099   0.009851798  16.576458664

Number of Observations: 2680
Number of Groups: 335
> VarCorr(model.e2)
prjId = pdLogChol(Time)
            Variance  StdDev   Corr
(Intercept) 20.962541 4.578487 (Intr)
Time        30.051822 5.481954 -0.567
Residual     7.689713 2.773033
```

Both Issues and PullReq have a positive impact on the Forks. But as time goes by, the impact of Issues on Forks will diminish.

VI Results & Discussions

6.1 Evaluations

Models	AIC	BIC
Model A	24518.55	24536.23
Model B	17339.53	17374.89
Model C1	16068.64	16115.79
Model C2	15826.99	15874.14
Model C3	15347.79	15394.94
Model C4	15710.01	15757.16
Model D1	15752.30	15811.24
Model D2	15221.89	15280.83
Model D3	15468.74	15527.67
Model D4	15151.71	15210.65
Model D5	15307.19	15366.13
Model D6	15174.78	15233.72
Model E1	**15132.34**	**15203.06**
Model E2	15290.82	15361.54

In summary, the results of Model E1 has the lowest value of AIC and BIC. Thus, Model E1 is the best model of all. From Model E1, we can conclude that Issues have a negative impact on the Forks. In the beginning, PullReq has a negative impact on the Forks, but over time, PullReq will have a positive impact on the Forks.

6.2 Limitations & Future Work

Although the correlation analysis of influential factors in the development process of GitHub open-source software has drawn many exciting conclusions, we only considered eight significant factors for the study and did not consider subsets of 8 significant factors, such as PullReqCmnt, IssuesCmnt and MemCommitters. Therefore, more influencing factors should be considered for further verification in future studies. Besides, the sample project selected for this article is only

335 open source projects, which is still a minimal number compared to GitHub's more than 12 million projects. Therefore, the results may have some contingency.

VII Conclusion

This paper analyzed the factors influencing the development of GitHub open-source software, proposed the effects of Issues, PullReq, CommitCmnt, and Commits on Forks under the time dimension, and analyzed the correlation among these factors. In future studies, we will consider more influencing factors and the correlation between multiple influencing factors with larger sample size.

References

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F. da V., … Vizcaíno, J. A. (2016). Ten Simple Rules for Taking Advantage of Git and GitHub. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004947

Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A Quick Introduction to Version Control with Git and GitHub. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004668

Kalliamvakou, E., Damian, D., Blincoe, K., Singer, L., & German, D. M. (2015). Open source-style collaborative development practices in commercial projects using GitHub. *Proceedings - International Conference on Software Engineering*. https://doi.org/10.1109/ICSE.2015.74

Dabbish, L., Stuart, C., Tsay, J., & Herbsleb, J. (2012). Social coding in GitHub: Transparency and collaboration in an open software repository. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. https://doi.org/10.1145/2145204.2145396

Tsay, J., Dabbish, L., & Herbsleb, J. (2014). Influence of social and technical factors for evaluating contribution in GitHub. *Proceedings - International Conference on Software Engineering*. https://doi.org/10.1145/2568225.2568315