Technical Report

Xunhui Zhang, Ayushi Rastogi, Yue Yu

March 9, 2020

This is the technical report for MSR 2020 data show case paper "On the Shoulders of Giants: A New Dataset for Pull-based Development Research".

1 Data Distribution

1.1 Dichotomous Metrics

Figure 1 shows the data distribution of dichotomous metrics.

- *same country* is the same_country metric. 81.7% contributor and integrator come from the same country.
- *same affiliation* is the same_affiliation metric. 90.4% contributor and integrator come from the same affiliation.
- *include test* is the test_inclusion metric. Only 19.5% pull requests include test code.
- *gender* is the contrib_gender metric. 90.2% contributors are male.
- *follow* is the contrib_follow_integrator metric. Only 7.13% contributors follow the closer of the pull request.
- *first pr* is the first_pr metric. 14.3% of the pull requests are submitted by contributors without any experience.
- *core* is the core_member metric. About 67.9% pull requests are submitted by core members.
- *conflict* is the comment_conflict metric. Only 1.19% pull requests' comments have "conflict" mark.
- *ci usage* is the ci_exists metric. 74.7% pull requests use CI tools.
- *ci pass* is the ci_test_passed metric. 69% of the pull requests passed the ci builds. 31% pull requests have 1 or more failures.
- *ci last status* is the ci_last_build_status metric. 87.9% pull requests passed the last build.
- *ci first status* is the ci_first_build_status metric. 75.5% pull requests passed the first build.

- *bug* is the bug_fix metric. 61.5% pull requests fix bugs, and 38.5% pull requests add new features.
- # is the hash_tag metric. 21.6% pull requests refer to other pull requests or issues.
- @ is the at_tag metric. 20.5% pull requests refer to developers.

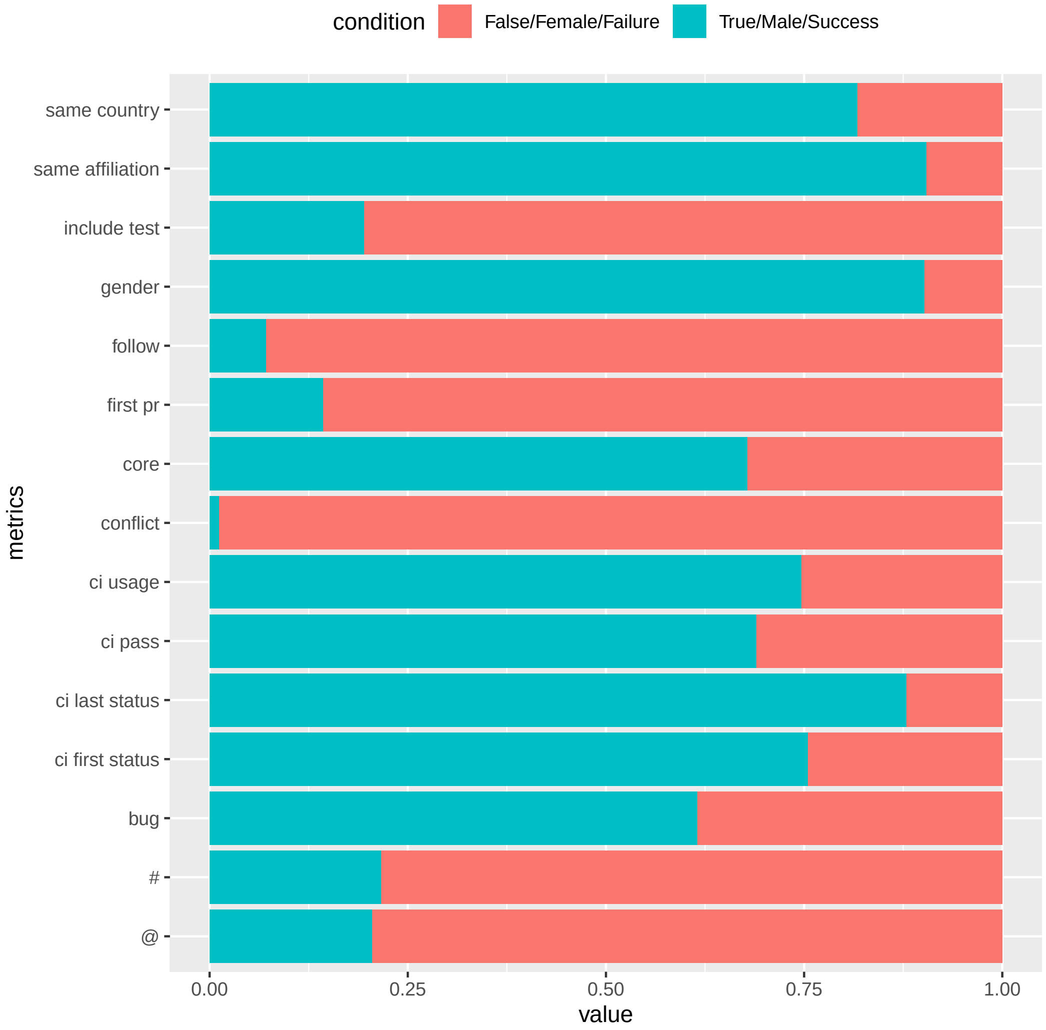


Figure 1: The distribution of dichotomous metrics

1.2 Continuous Metrics

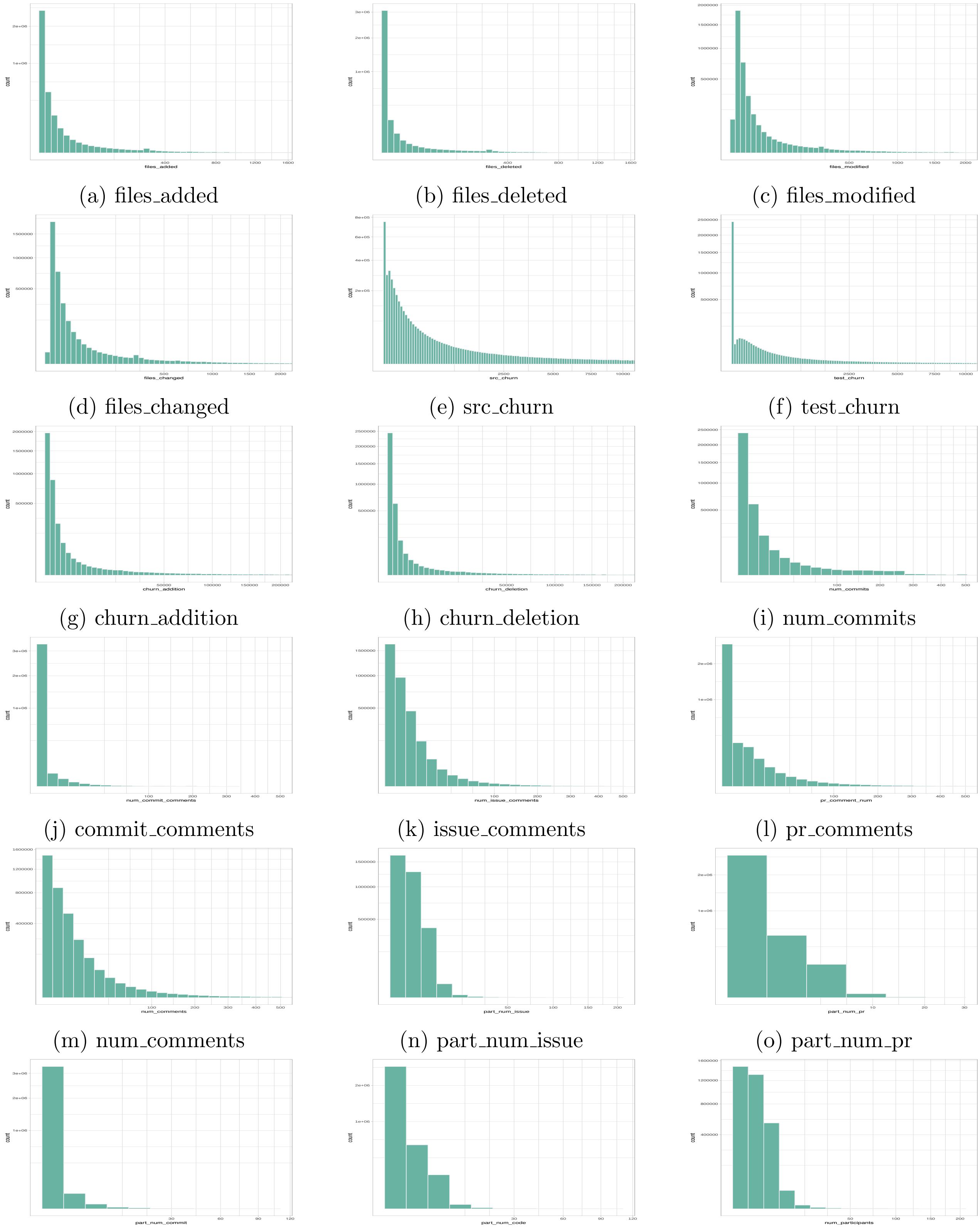
Figure 2, 3, 4, 5 show the data distribution of continuous metrics with square root scale. For some of the metrics, we did some pre-processings.

- *num_commits*: we add a range of x axis from 0 to 500, and we only consider those pull request that have greater or equal to 1 commit.
- *src_churn*: we add a range of x axis from 0 to 10,000.

- *test_churn*: we add a range of x axis from 0 to 10,000.
- *files_added*: we add a range of x axis from 0 to 1,500.
- *files_deleted*: we add a range of x axis from 0 to 1,500.
- *files_modified*: we add a range of x axis from 0 to 2,000.
- *files_changed*: we add a range of x axis from 0 to 2,000.
- *src_files*: we add a range of x axis from 0 to 2,000.
- *doc_files*: we add a range of x axis from 0 to 2,000.
- *other_files*: we add a range of x axis from 0 to 2,000.
- *num_commit_comments*: we add a range of x axis from 0 to 500.
- *num_issue_comments*: we add a range of x axis from 0 to 500.
- *num_pr_comments*: we add a range of x axis from 0 to 500.
- *num_comments*: we add a range of x axis from 0 to 500.
- *churn_addition*: we add a range of x axis from 0 to 200,000.
- *churn_deletion*: we add a range of x axis from 0 to 200,000.
- *ci_latency*: we add a range of x axis from 0 to 10,000,000.
- *ci_failed_perc*: we only consider those pull requests that use ci tools.
- *pr_succ_rate*: we only consider contributors who had submitted pull requests before.
- *perc_neg_emotion*: we only consider pull requests that have at least 1 comment.
- *perc_pos_emotion*: we only consider pull requests that have at least 1 comment.
- *perc_neu_emotion*: we only consider pull requests that have at least 1 comment.
- *perc_contrib_neg_emo*: we only consider pull requests that have at least 1 comment.
- *perc_contrib_pos_emo*: we only consider pull requests that have at least 1 comment.
- *perc_contrib_neu_emo*: we only consider pull requests that have at least 1 comment.
- *perc_inte_neg_emo*: we only consider pull requests that have at least 1 comment.
- *perc_inte_pos_emo*: we only consider pull requests that have at least 1 comment.
- *perc_inte_neu_emo*: we only consider pull requests that have at least 1 comment.

1.3 Factor Metrics

Figure 6 shows the data distribution of factor metrics. For *contrib_country*, *inte_country*, *contrib_affiliation* and *inte_affiliation*, we show the top 6 factors, and treat other factors as others.



(a) files_added

(b) files_deleted

(c) files_modified

(d) files_changed

(e) src_churn

(f) test_churn

(g) churn_addition

(h) churn_deletion

(i) num_commits

(j) commit_comments

(k) issue_comments

(l) pr_comments

(m) num_comments

(n) part_num_issue

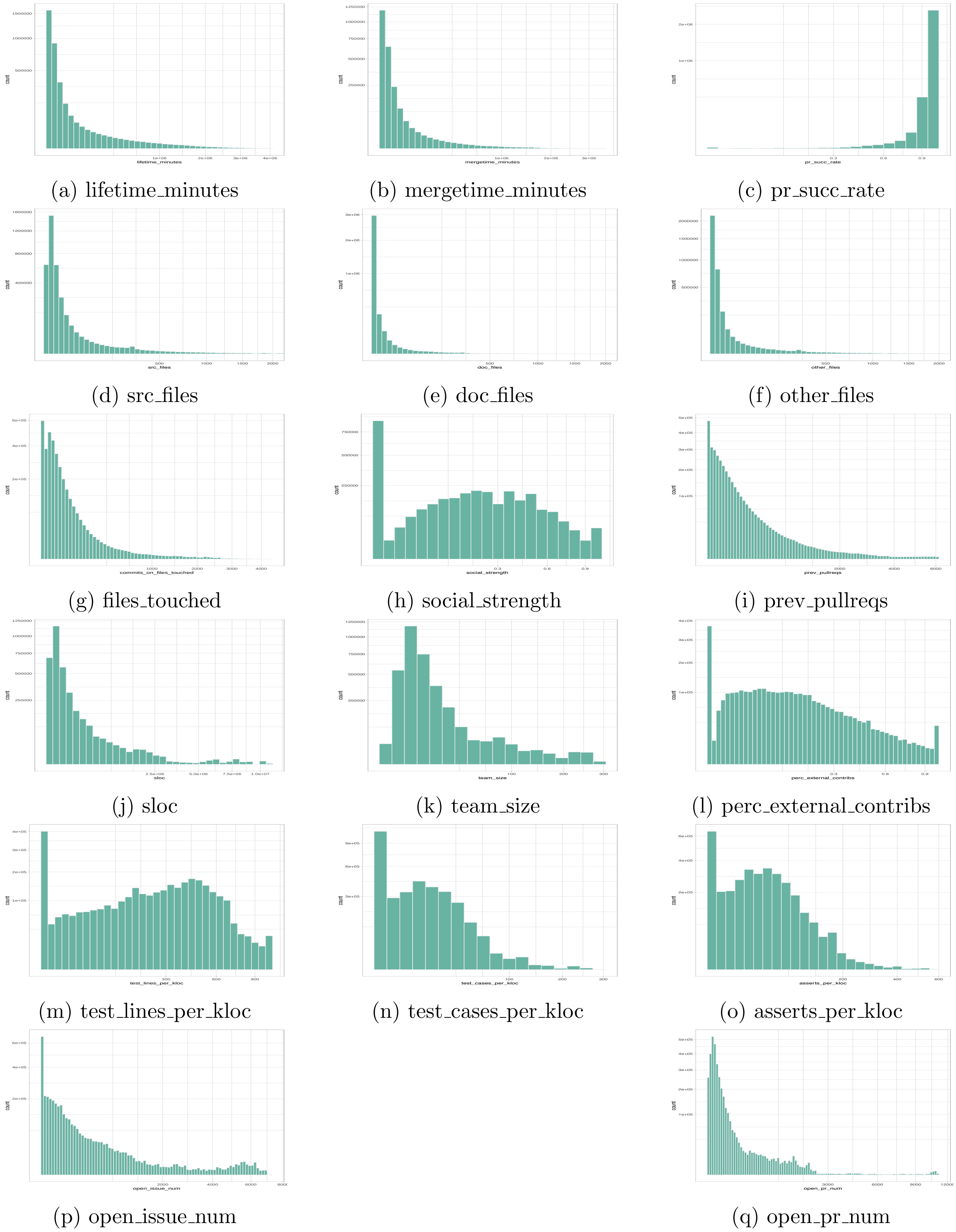
(o) part_num_pr

(p) part_num_commit

(q) part_num_code

(r) num_participants

Figure 2: The distribution of continuous metrics



(a) lifetime_minutes

(b) mergetime_minutes

(c) pr_succ_rate

(d) src_files

(e) doc_files

(f) other_files

(g) files_touched

(h) social_strength

(i) prev_pullreqs

(j) sloc

(k) team_size

(l) perc_external_contribs

(m) test_lines_per_kloc

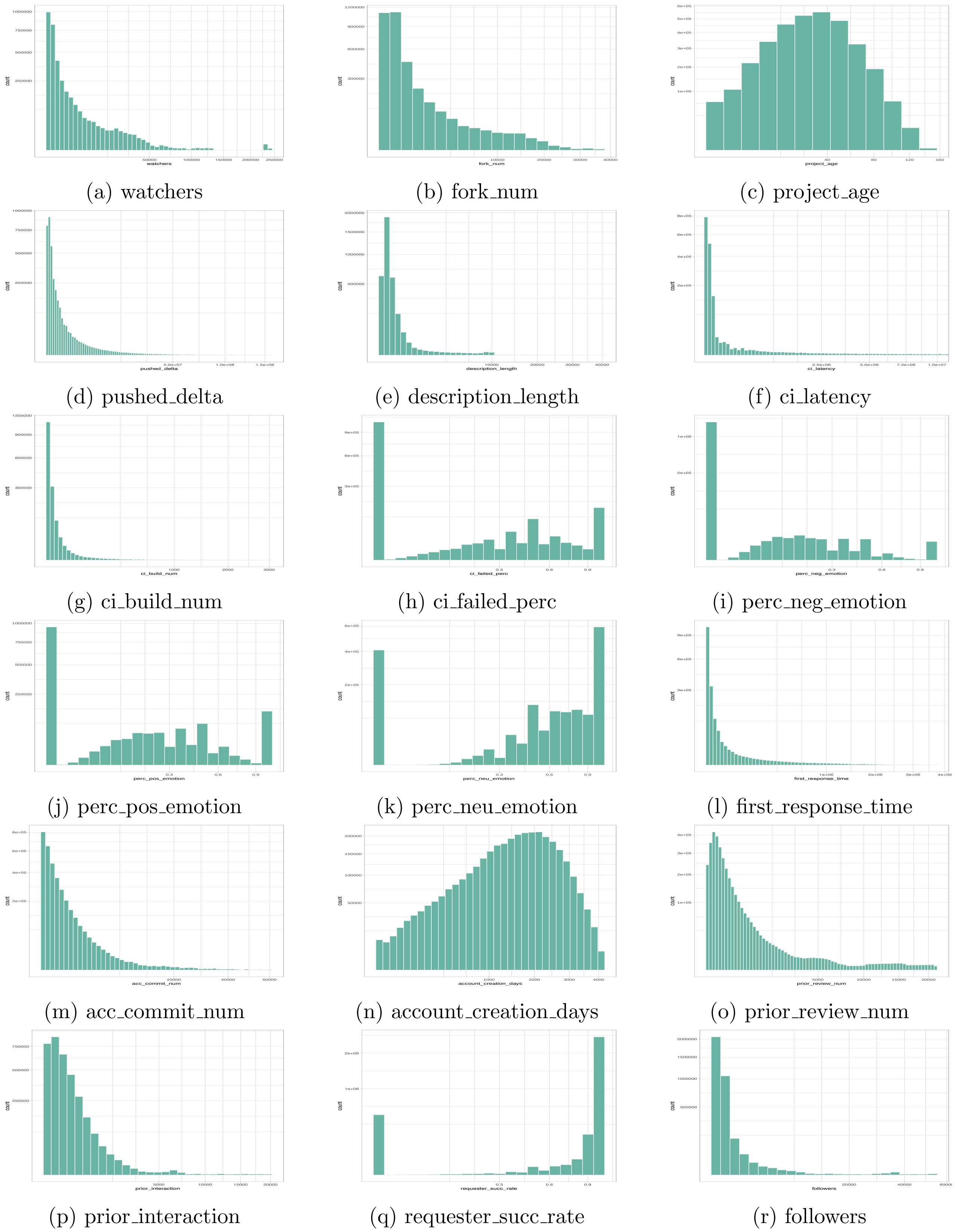
(n) test_cases_per_kloc

(o) asserts_per_kloc

(p) open_issue_num

(q) open_pr_num

Figure 3: The distribution of continuous metrics



(a) watchers

(b) fork_num

(c) project_age

(d) pushed_delta

(e) description_length

(f) ci_latency

(g) ci_build_num

(h) ci_failed_perc

(i) perc_neg_emotion

(j) perc_pos_emotion

(k) perc_neu_emotion

(l) first_response_time

(m) acc_commit_num

(n) account_creation_days

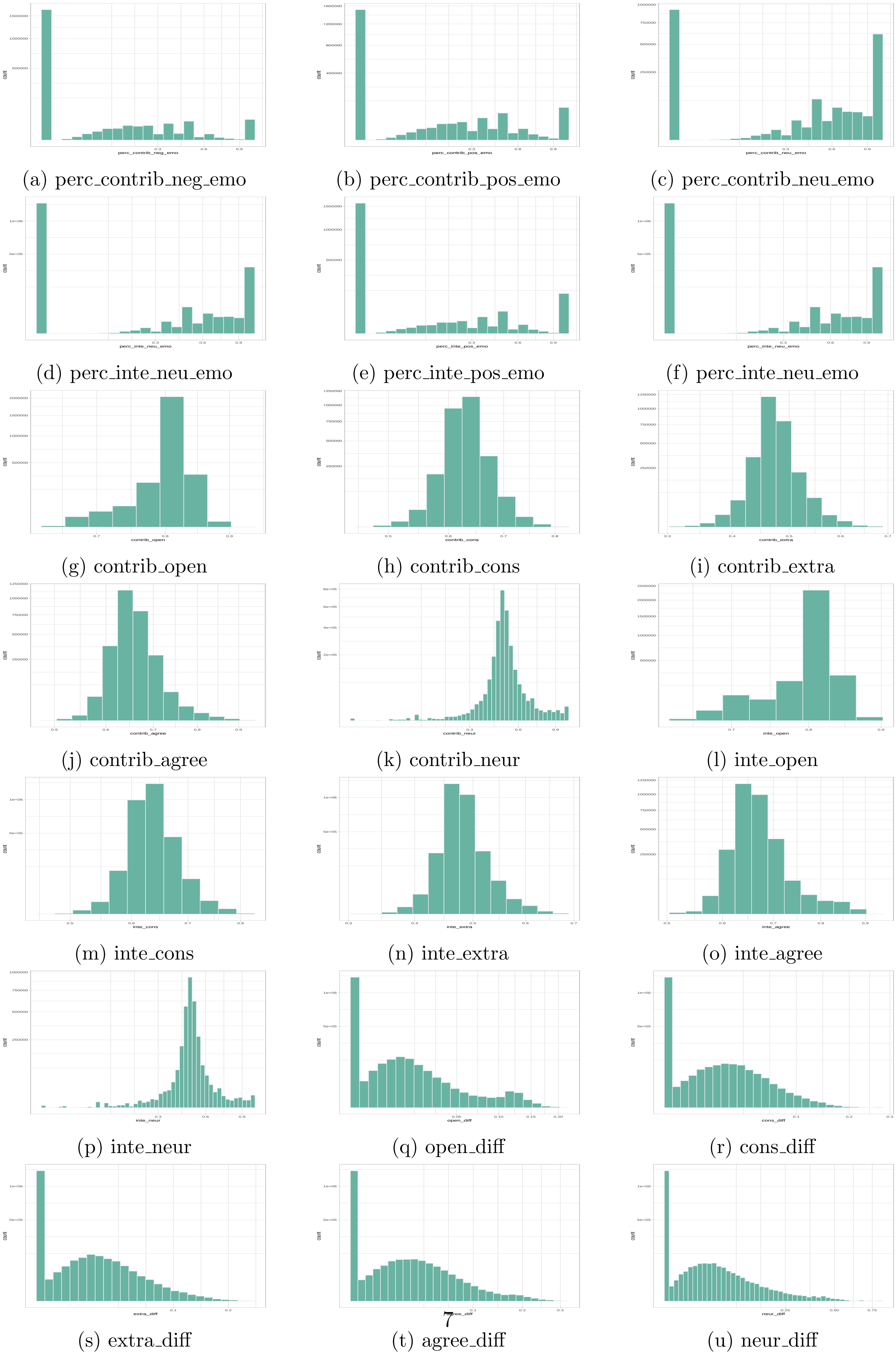
(o) prior_review_num

(p) prior_interaction

(q) requester_succ_rate

(r) followers

Figure 4: The distribution of continuous metrics



(a) perc_contrib_neg_emo

(b) perc_contrib_pos_emo

(c) perc_contrib_neu_emo

(d) perc_inte_neu_emo

(e) perc_inte_pos_emo

(f) perc_inte_neu_emo

(g) contrib_open

(h) contrib_cons

(i) contrib_extra

(j) contrib_agree

(k) contrib_neur

(l) inte_open

(m) inte_cons

(n) inte_extra

(o) inte_agree

(p) inte_neur

(q) open_diff

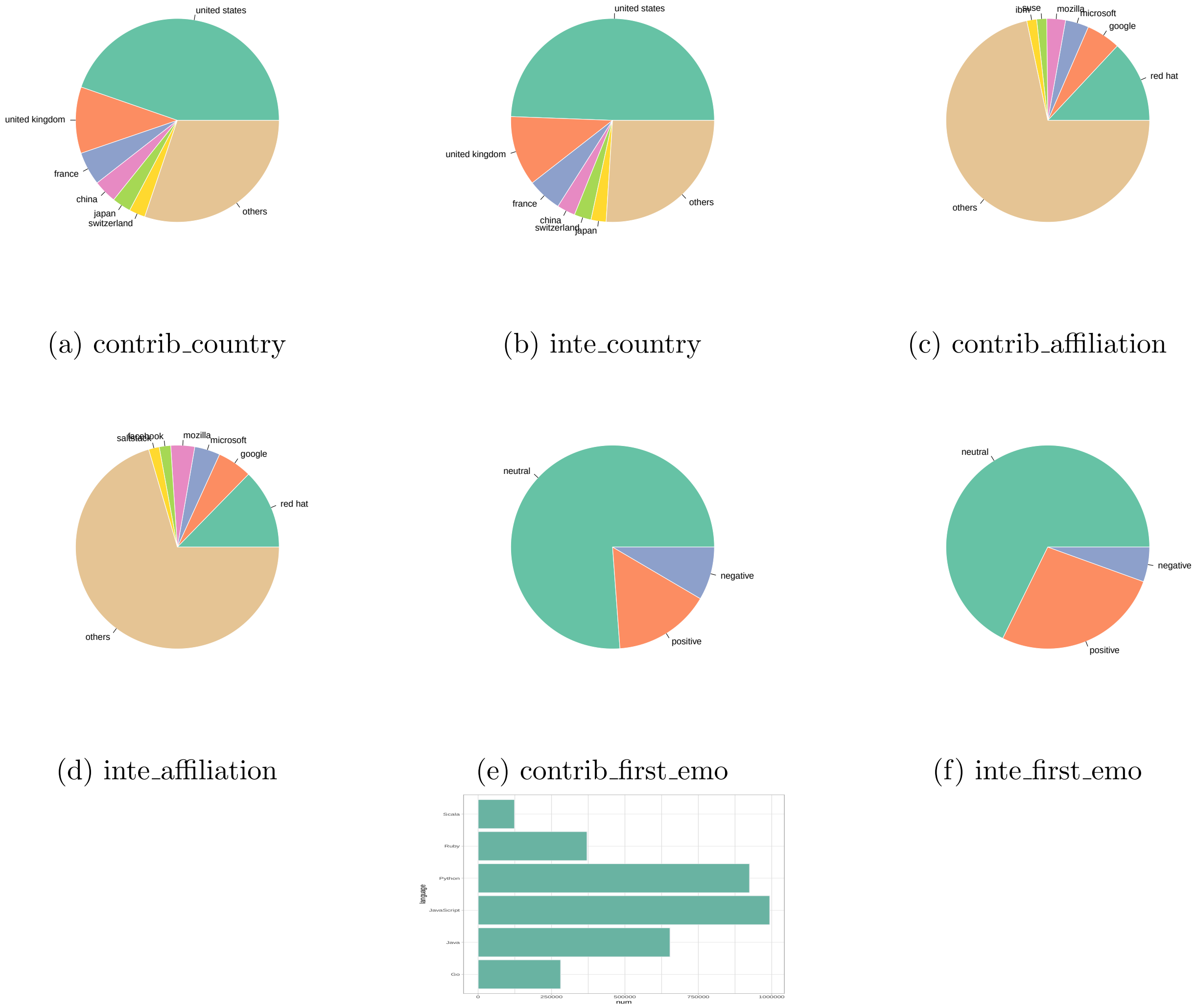
(r) cons_diff

(s) extra_diff

(t) agree_diff

(u) neur_diff

Figure 5: The distribution of continuous metrics



(a) contrib_country

(b) inte_country

(c) contrib_affiliation

(d) inte_affiliation

(e) contrib_first_emo

(f) inte_first_emo

(g) language

Figure 6: The distribution of factor metrics