Technical Report

Xunhui Zhang, Ayushi Rastogi, Yue Yu

March 12, 2020

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

1 Data Distribution

1.1 Categorical Metrics

1.1.1 Binary Metrics

Figure 1 shows the data distribution of binary metrics, and Table 1 presents the proportion of each level.

Table 1: Proportion of each binary categorical feature

Feature	Proportion	Feature	Proportion
same_country	True(81.7%); False(18.3%)	same_affiliation	True(90.4%); False(9.6%)
contrib_gender	Male(90.2%); Female(9.8%)	test_inclusion	True(19.5%); False(80.5%)
contrib_follow_integrator	True(7.1%); False(92.9%)	first_pr	True(14.3%); False(85.7%)
comment_conflict	True(1.2%); False(98.8%)	core_member	True(67.9%); False(32.1%)
ci_test_passed	True(69%); False(31%)	ci_exists	True(74.7%); False(25.3%)
ci_first_build_status	Success(75.5%); Failure(24.5%)	bug_fix	True(61.5%); False(38.5%)
ci_last_build_status	Success(87.9%); Failure(12.1%)	hash_tag	True(21.6%); False(78.4%)
at_tag	True(20.5%); False(79.5%)		

1.1.2 Multi-level Metrics

Figure 2 shows the data distribution of multi-level categorical metrics. For *contrib_country*, *inte_country*, *contrib_affiliation* and *inte_affiliation*, we show the top 6 factors, and treat other factors as *others*. Table 2 shows the proportion of each level.

1.2 Continuous Metrics

Figure 3, 4, 5, 6 show the data distribution of continuous metrics with square root scale.

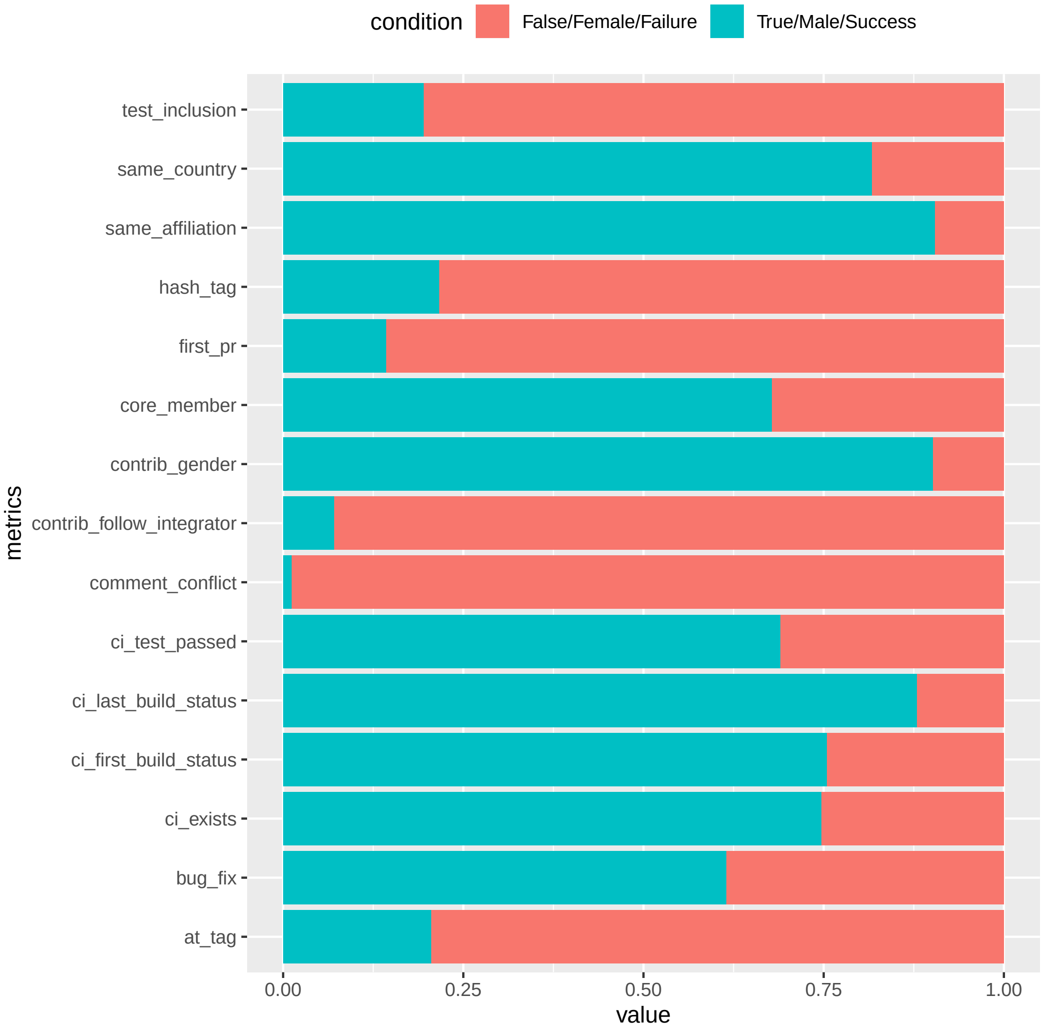
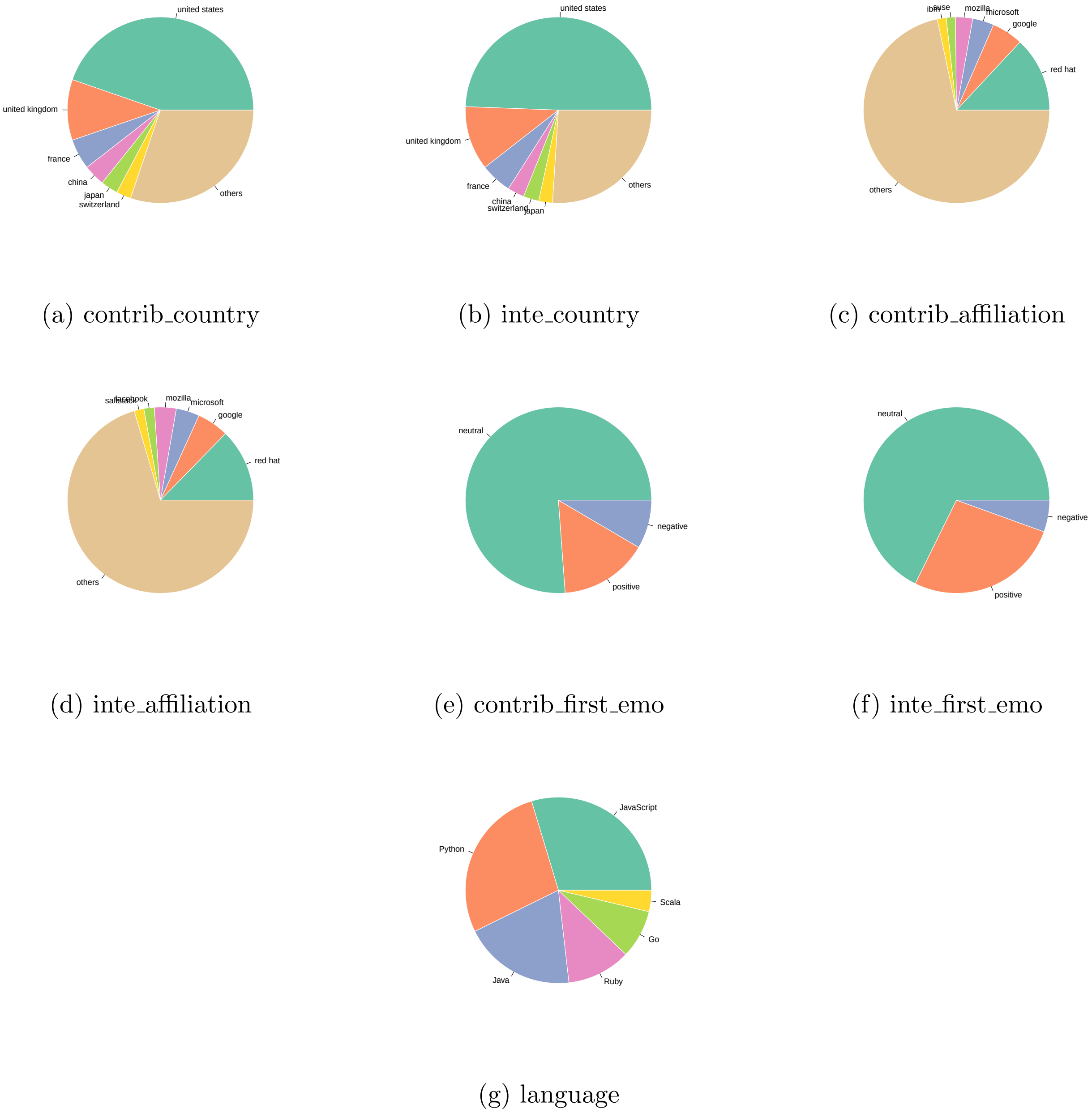


Figure 1: The distribution of dichotomous metrics

Table 2: Proportion of each multi-level categorical feature

Feature	Proportion
contrib_country	US(44.7%); UK(10.6%); France(5.3%); China(3.7%); Japan(3.0%); Switzerland(2.6%); others(30.1%)
inte_country	US(49.4%); UK(11.1%); France(5.5%); China(2.9%); Switzerland(2.7%); Japan(2.4%); others(26.0%)
contrib_affiliation	red hat(13.2%); Google(5.5%); Microsoft(3.7%); Mozilla(3.0%); SUSE(1.6%); IBM(1.6%); others(71.4%)
inte_affiliation	red hat(12.8%); Google(5.6%); Microsoft(4.1%); Mozilla(3.8%); Facebook(1.8%); SaltStack(1.7%); others(70.2%)
contrib_first_emo	negative(8.5%); positive(15.4%); neutral(76.1%)
inte_first_emo	negative(5.5%); positive(26.8%); neutral(67.7%)
language	JavaScript(29.7%); Python(27.6%); Java(19.5%); Ruby(11.1%); Go(8.4%); Scala(3.7%)



(a) contrib_country

(b) inte_country

(c) contrib_affiliation

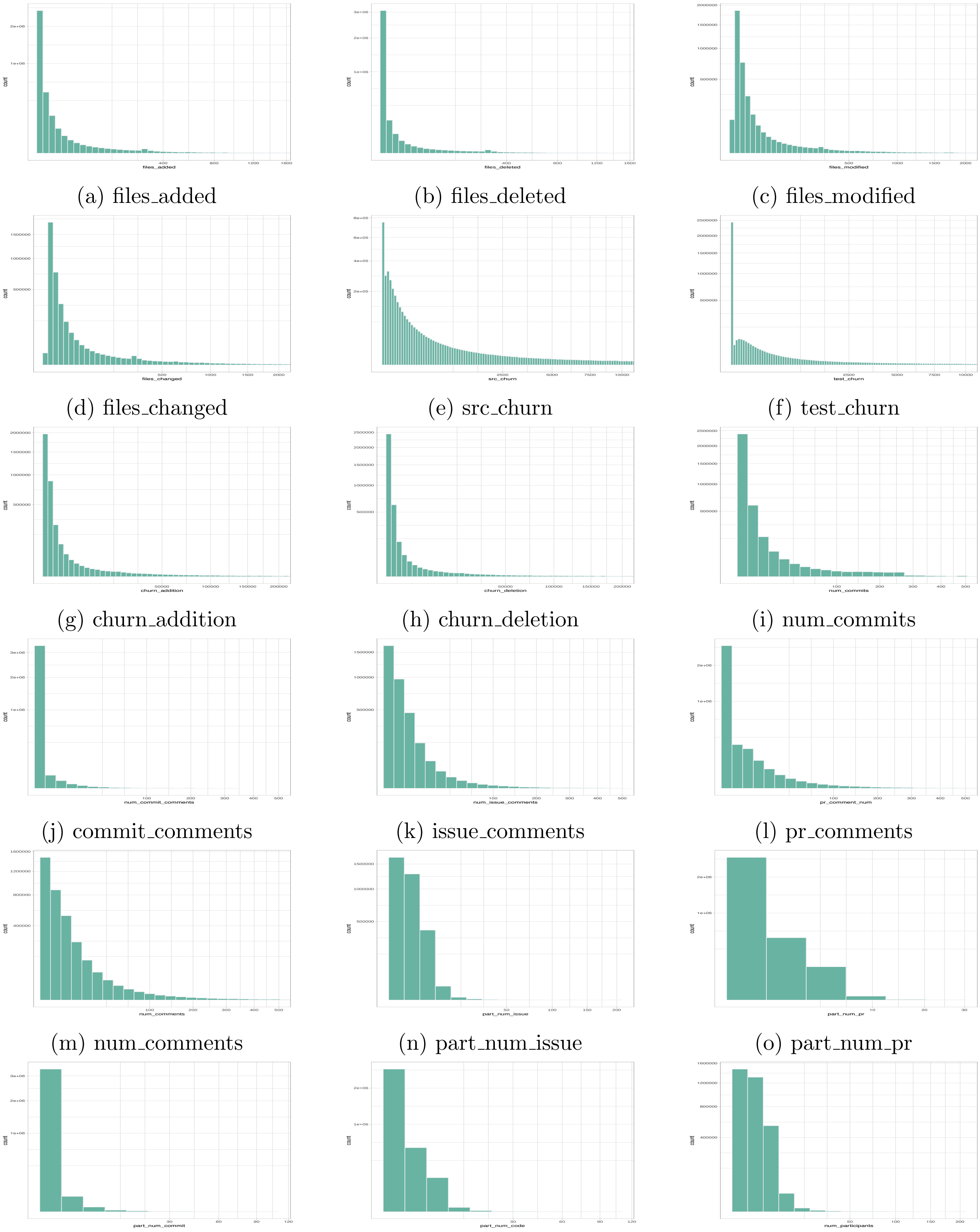
(d) inte_affiliation

(e) contrib_first_emo

(f) inte_first_emo

(g) language

Figure 2: The distribution of multi-level categorical metrics



(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

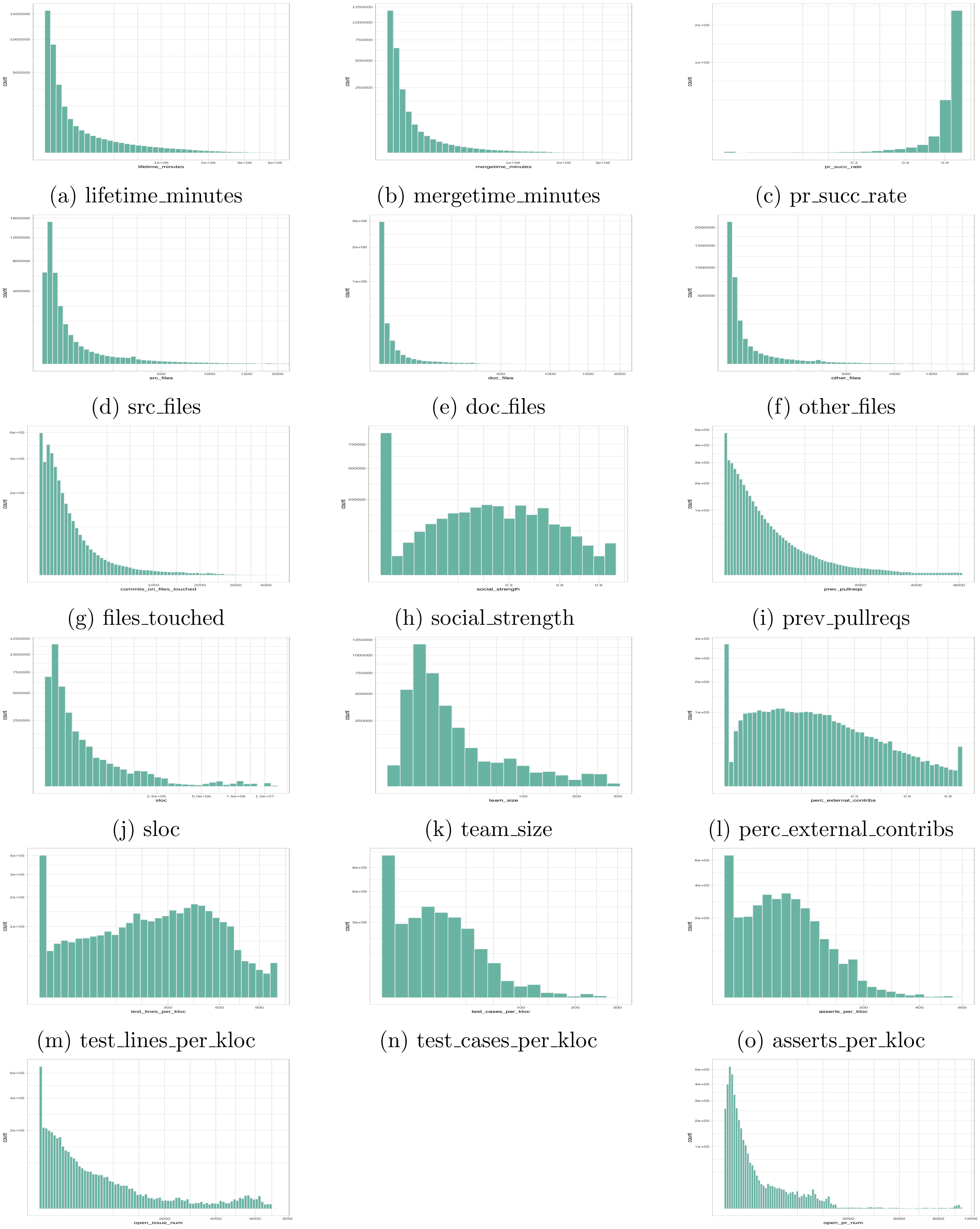
(o) part_num_pr

(p) part_num_commit

(q) part_num_code

(r) num_participants

Figure 3: 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

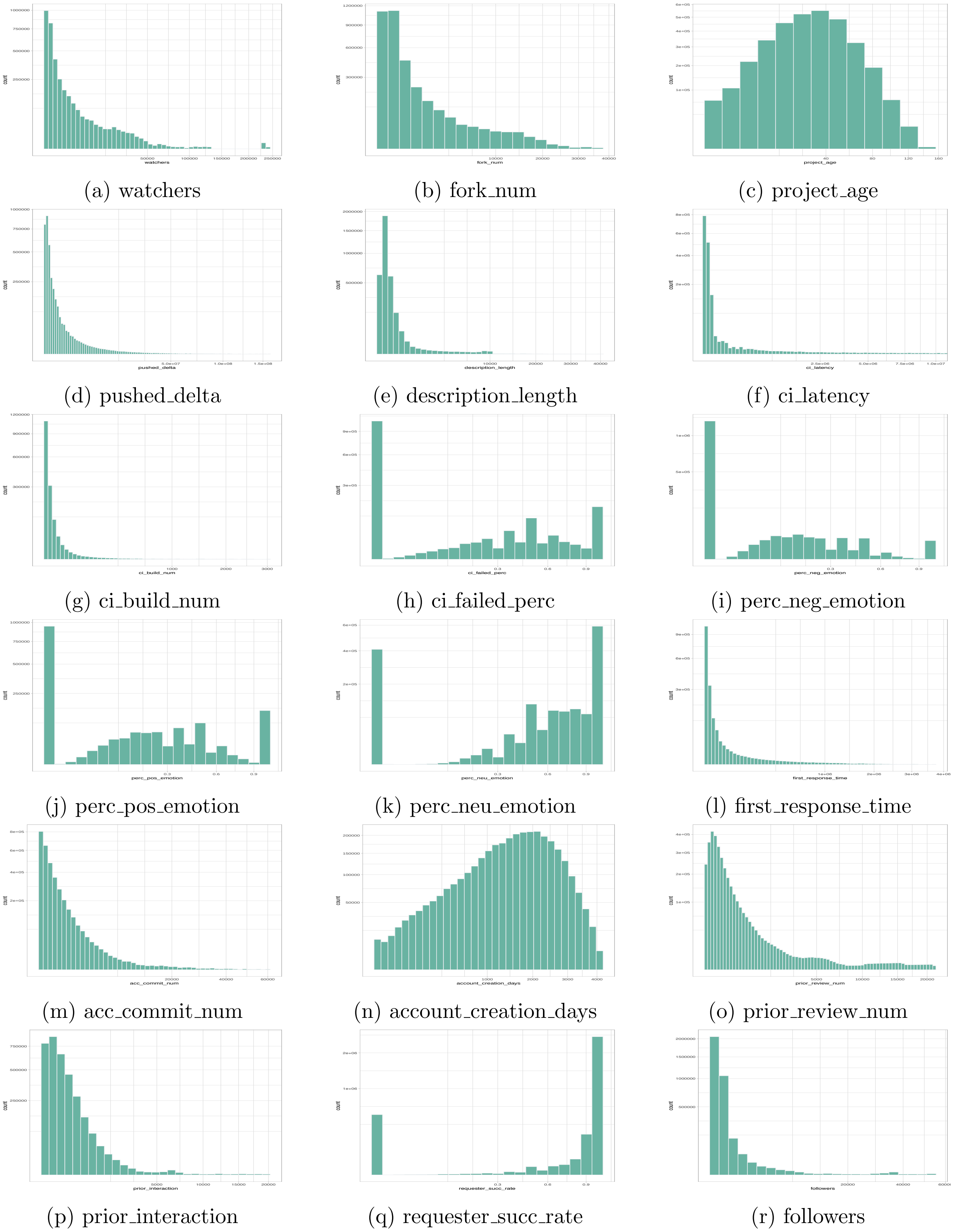
(n) test_cases_per_kloc

(o) asserts_per_kloc

(p) open_issue_num

(q) open_pr_num

Figure 4: 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

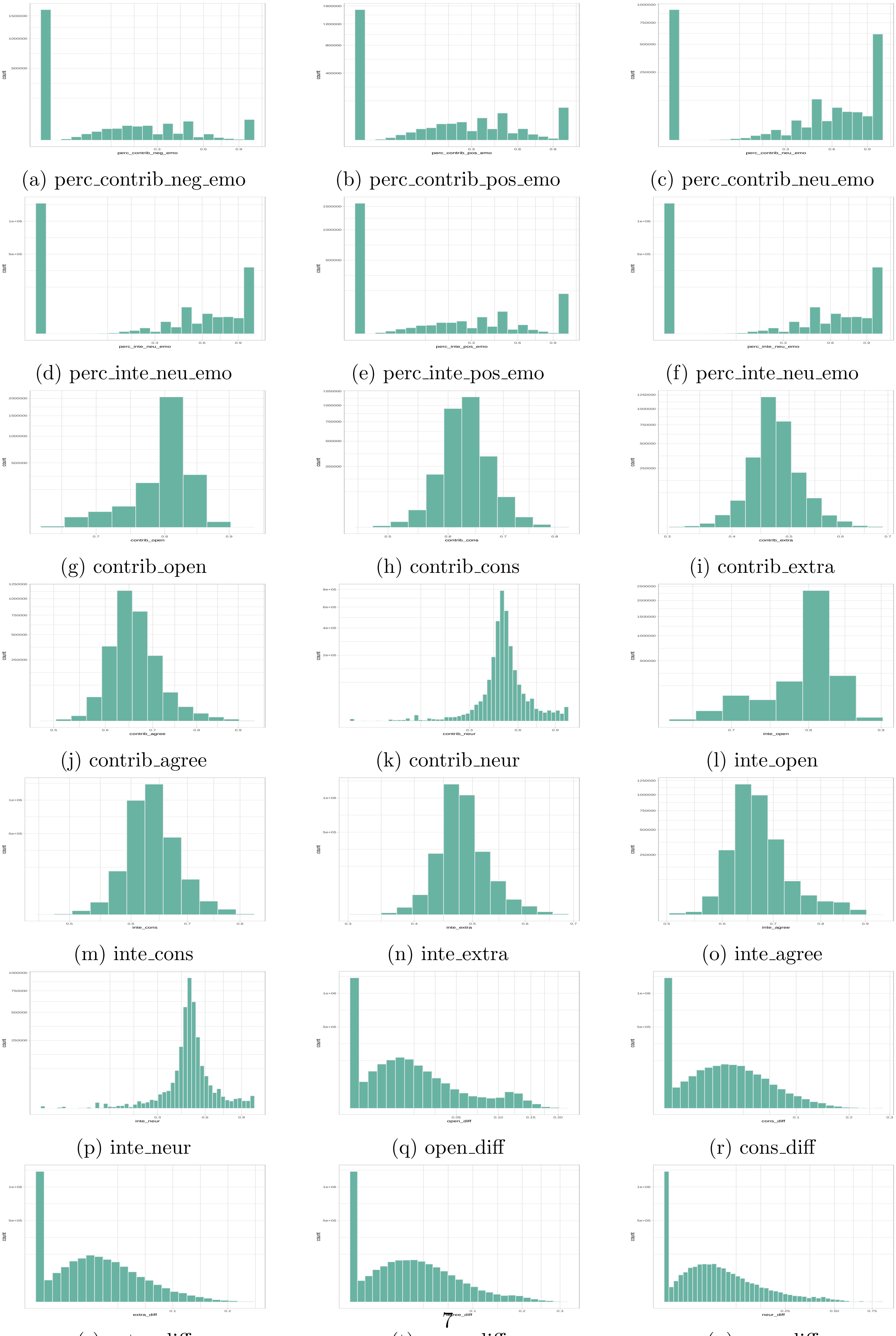
(o) prior_review_num

(p) prior_interaction

(q) requester_succ_rate

(r) followers

Figure 5: 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 6: The distribution of continuous metrics