

How Developers Use Relation Chains in Gerrit-Based Review Ecosystems: An Empirical Study Across Three Open-Source Ecosystems

Ahmed Belhouchette ✉

ENSI, Manouba University, Tunisia

Moataz Chouchen ✉

Concordia University, Montreal, QC, Canada

Marouene Chaieb ✉

ENSI, Manouba University, Tunisia

LARODEC Laboratory, ISG-Tunis, Tunis University, Tunisia

Mohammad Hamdaqa ✉

Polytechnique Montréal, Montreal, QC, Canada

Abdelwahab Hamou-Lhadj ✉

Concordia University, Montreal, QC, Canada

Abstract

Background. Developers increasingly coordinate dependent review workflows by submitting sequences of related changes rather than monolithic ones. In Gerrit, these dependencies form *relation chains*: structured review units that link changes together. As chains become more common, they shape review activities through synchronization overhead, CI amplification, and merge-ordering constraints.

Aim. We investigate how developers adopt relation chains and how these dependency structures influence review dynamics and outcomes.

Method. We analyze 29,580 relation chains from 15 repositories across three Gerrit ecosystems (OpenStack, Wikimedia, ONAP), comprising 401,256 changes, using Mann–Kendall trend tests, Mann–Whitney with Cliff’s δ for chain-vs-solo comparison, and Spearman correlations for base–descendant dependency.

Results. Chain prevalence ranges from 5% to 49% across projects, increasing in 14 of 15. Chain changes take a median of $2.6\times$ longer to merge than size-matched solo changes, with the gap widening for very large changes. Review effort propagates through dependency-linked review workflows: base-change review activity co-varies with descendant review activity ($\rho = 0.43\text{--}0.61$ in 14–15/15 projects), and 33.5% of chain members undergo structural evolution during review.

Conclusions. Relation chains operate as durable, ecosystem-shaped coordination units with internal structure that change-centric analyses cannot capture. Future review analytics, reviewer-assignment systems, and AI-assisted review tools should reason over chains rather than isolated changes.

2012 ACM Subject Classification Software and its engineering → Software configuration management and version control systems; Software and its engineering → Collaboration in software development; Software and its engineering → Software verification and validation

Keywords and phrases Code review, relation chains, stacked changes, dependent patches, empirical software engineering, Gerrit

Digital Object Identifier 10.4230/LIPIcs.ESEM.2026.26

Category Technical Track Paper



© Ahmed Belhouchette, Moataz Chouchen, Marouene Chaieb, Mohammad Hamdaqa, and Abdelwahab Hamou-Lhadj;

licensed under Creative Commons License CC-BY 4.0

20th International Symposium on Empirical Software Engineering and Measurement (ESEM 2026).

Editors: Robert Feldt, Maria Paasivaara, Daniel Mendez, Stefan Wagner, and Marvin Muñoz Barón;

Article No. 26; pp. 26:1–26:21



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Modern Code Review (MCR) is used to detect defects, discuss design decisions, transfer knowledge, and enforce coding conventions [6, 34, 29, 24, 26]. Large engineering tasks such as refactorings, migrations, and multi-component features are increasingly decomposed into sequences of dependent changes submitted as *stacked changes*. In Gerrit, these structures are represented as *relation chains*, a higher-level coordination abstraction in modern code-review workflows; similar abstractions exist in Phabricator, Graphite, GitHub, and **ghstack** [32, 20, 19, 30].

Following the Gerrit documentation [18], we use the term *relation chain* to denote a dependency-linked review structure in which multiple changes evolve simultaneously and the state of one review directly influences the validation, synchronization, and mergeability of dependent reviews. In this paper, relation chains refer specifically to Gerrit parent-SHA dependency structures reconstructed from patchset-level parent relationships; while similar workflows exist in GitHub stacked pull requests, Phabricator, and **ghstack**, our operationalization and findings are specific to Gerrit-based review ecosystems. Relation chains let reviewers reason about large engineering efforts as coordinated structures rather than isolated changes. Figure 1 shows an OpenStack Neutron example: a **pylint** migration¹ decomposed into descendants addressing different warning categories. The same dependency that enables this coordination can also become a source of friction: updates to one chain member can cascade across descendants through forced rebasing, repeated CI validation, and delayed merge propagation. On ONAP oom change 100114², a reviewer warned the author that “*as parent is abandoned, we cannot merge... please rebase on master*,” surfacing parent abandonment, merge blocking, and forced rebasing as everyday chain-coordination costs. Understanding how such dependency structures influence review processes matters for review analytics, prioritization, and tooling [39].

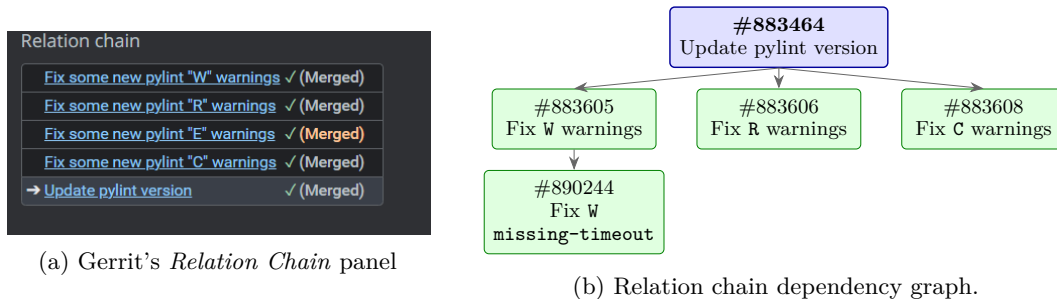


Figure 1 A branching relation chain in OpenStack Neutron (change #883464¹). (a) Gerrit’s flat *Relation Chain* panel. (b) The same chain reconstructed from patchset-level parent SHAs: blue is the base, green is a descendant. All five members merged.

Despite their growing adoption, prior research has overlooked relation chains as first-class review artifacts. Existing work captures only fragments: studies of cross-component dependencies [3] model pairwise relationships between repositories rather than the chain that links a sequence of changes within a single review; **Depends-On/Needed-By** annotation analyses [2] require the dependency to be declared in the commit message, missing chains

¹ <https://review.opendev.org/c/openstack/neutron/+883464>

² <https://gerrit.onap.org/r/c/oom/+100114>

created implicitly through parent-SHA links; and patch-linkage research [23, 44, 43] traces references between review discussions but not the structural dependency graph that determines merge order. Properties that emerge only when a change is read together with its in-review parents and children (friction propagation from base to descendant, mid-review restructuring of the dependency graph, position-dependent review effort) fall outside what existing methods can observe. These workflow effects remain largely invisible when reviews are analyzed as isolated artifacts. The literature has no empirical characterisation of relation chains as the unit through which developers coordinate dependent review work at scale.

To address this gap, we present an empirical characterisation of relation chains as first-class review artifacts. We analyze 401,256 reviewed changes and 29,580 chains across three Gerrit ecosystems (OpenStack, Wikimedia, ONAP), selected to span different development cultures, CI-gating regimes, and community sizes. The study combines three lenses: temporal-trend analysis of chain adoption, stratified comparison of chain members against size-matched solo changes, and dependency-aware correlation analysis between base changes and their descendants. Together they let us characterise chains both as a structural feature of the review graph and as a coordination mechanism whose dynamics shape review outcomes that change-level studies cannot observe. Because our analyses are observational, we interpret the reported effects as associations rather than causal consequences of chain membership.

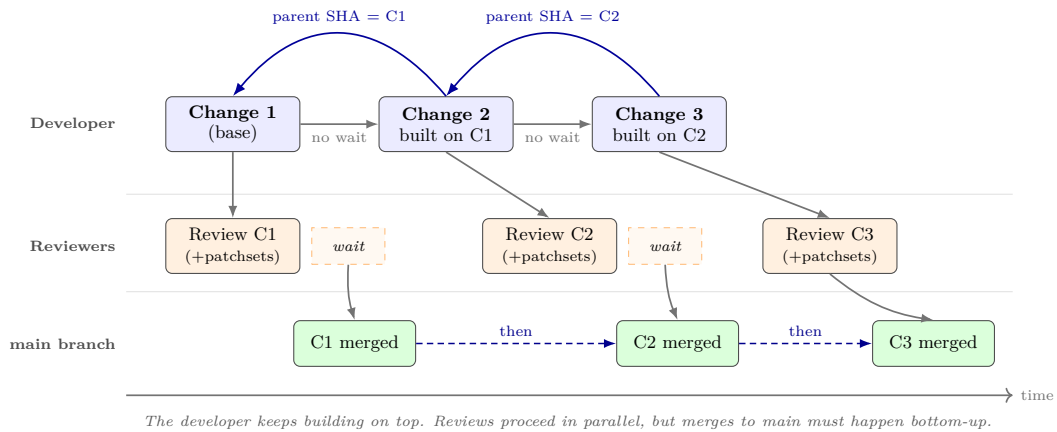
Our results converge on a single characterisation: relation chains are durable, ecosystem-shaped, internally-structured coordination units whose review dynamics differ from those of isolated changes. Each of the four findings below contributes evidence along one of these dimensions. First, chains are *prevalent and growing*: prevalence ranges from 5% to 49% and increases in 10 of 15 projects, so dependent-change workflows are not a niche pattern. Second, chains carry *review costs not visible at the change level*: members take a median of $2.6\times$ longer to merge than size-matched solo changes, with overhead concentrating in very large changes (Cliff's $\delta = 0.49$), and CI activity is amplified five- to twenty-fold under strict gating. Third, chains have *internal structure that shapes review outcomes*: middle-of-chain members absorb the highest review effort, and base-change review activity co-varies with descendant activity ($\rho = 0.43\text{--}0.61$ in 14–15 of 15 projects), a regularity we call the *foundation effect*. Fourth, chains are *not static*: 33.5% of members undergo a structural change in their parent SHA before merging, and chains can persist across multi-year intermittent timespans, with the longest merge gap reaching 2.85 years on a chain in OpenStack *Neutron*. These results call into question the default assumption underlying current review tooling, which processes each change independently [40, 31, 25], and motivate chain-aware review analytics, reviewer-assignment systems that prioritise members whose merge unblocks the largest downstream subgraph, and AI-assisted review agents that condition on chain position and base-change review state.

2 Background: Modern Code Review, Gerrit, and Relation Chains

Modern Code Review (MCR) is a tool-supported form of peer review used in industrial and open-source development [6, 34, 26], supporting defect detection, design discussion, knowledge transfer, and coding-convention enforcement [29, 24, 8]. Smaller, focused changes are easier to review and more likely to be accepted [45, 33]; larger engineering tasks therefore motivate developers to split work across related changes.

Gerrit is a web-based code review platform used by projects such as OpenStack, Android, Chromium, Wikimedia, and ONAP [17, 39]. Each review request is a *change* identified by a *Change-Id* and refined through successive *patchsets*. When a developer uploads a change whose parent commit corresponds to another change still under review, Gerrit links them as a parent-child dependency; a sequence of such linked changes forms a *relation chain* [18].

Relation chains let developers continue dependent work without waiting for earlier changes to merge. The same workflow appears in Phabricator [32], Graphite [20], Meta’s **ghstack** [30], and GitHub’s native stacked PRs [19]; earlier work on Linux kernel patch series [35] shows dependent review workflows predate modern platforms. Figure 2 illustrates a typical workflow. Throughout this paper, *base*, *middle*, and *top* denote the first, intermediate, and last members of a chain, and *solo* denotes a change with no in-review parent or child.



■ **Figure 2** Working with a relation chain in Gerrit. The developer builds C2 on C1 and C3 on C2 without waiting. Reviews proceed in parallel; merges into the main branch are bottom-up.

3 Related Work

We position our study with respect to three lines of work: (i) empirical studies of modern code review that examine review outcomes at the change level, (ii) reviewer recommendation and review automation systems that consume change-level features, and (iii) change decomposition and dependent-review research that touches on multi-change workflows. For each, we summarise what the line has shown and clarify how our chain-level focus differs.

3.1 Empirical Studies in Modern Code Review

Empirical work has examined MCR at the change level, characterizing factors that shape review latency and outcomes [8, 24, 47, 48] and the role of reviewer expectations and practice [6, 34, 26]. McIntosh et al. [29] showed that low review coverage and participation correlate with post-release defects; follow-up work examined review participation across Android, Qt, and OpenStack [39] and code ownership in MCR [38]. Change size is a recurring predictor [45, 33], and CI plays a parallel role: prior studies analyzed its effect on review duration [9], the cost of repeated builds in OpenStack [27], and bot activity in review [46].

Key difference from our work. Prior empirical MCR studies treat each review as an independent unit. We shift the focus to chain-level coordination (review effort, conversation, and dependency-graph evolution when a change belongs to a relation chain) and identify intra-chain structural factors such as chain position, base-change activity, and structural evolution that change-level analyses cannot observe.

3.2 Reviewer Recommendation and Review Automation

Two adjacent lines of work aim to reduce review effort through tooling. *Reviewer recommendation* systems suggest reviewers based on file-location history [40] or expertise and workload [31], while pull-request decision factors have been studied on GitHub [48]. *LLM-based review automation* extends this with pre-trained models that generate or refine review comments [25, 41, 21]. Related work has built predictive models on change-level features for merged-vs-abandoned classification [11], completion-time prediction [12], and effort-aware prioritization [10].

Key difference from our work. Existing tools operate on individual changes and do not consume chain-level signals such as chain position, base-change review state, or evolving dependency graphs. Our study shows the importance of accounting for this dependency context when reasoning about review outcomes, providing a concrete extension point for chain-aware tooling.

3.3 Change Decomposition and Dependent Reviews

The benefits of smaller review units create a tension: focused changes improve reviewability, but real engineering tasks are often too large for a single review. Prior work has studied *change decomposition*: Barnett et al. [7] proposed automated decomposition for tangled changesets at Microsoft, and Di Biase et al. [15] showed decomposed changes reduce false-positive review comments and improve review quality. Related work on dependencies: Arabat and Sayagh [3] found inter-component dependencies persist in microservice systems; Arabat et al. [2] examined explicit *Depends-On/Needed-By* annotations in OpenStack; *patch linkages* [23, 44, 43] trace references between review discussions; and *change coupling* [16, 49] mines version histories for co-evolving files.

Key difference from our work. Decomposition research treats decomposition as a property of individual changes; dependency research models dependencies as pairwise or textual. Relation chains are the mechanism by which developers implement decomposition in practice, introducing structural dynamics (base-change effects on descendants, merge-ordering constraints, dependency-graph evolution) that change-level studies cannot observe.

4 Study Design

Figure 3 shows an illustration of our study.

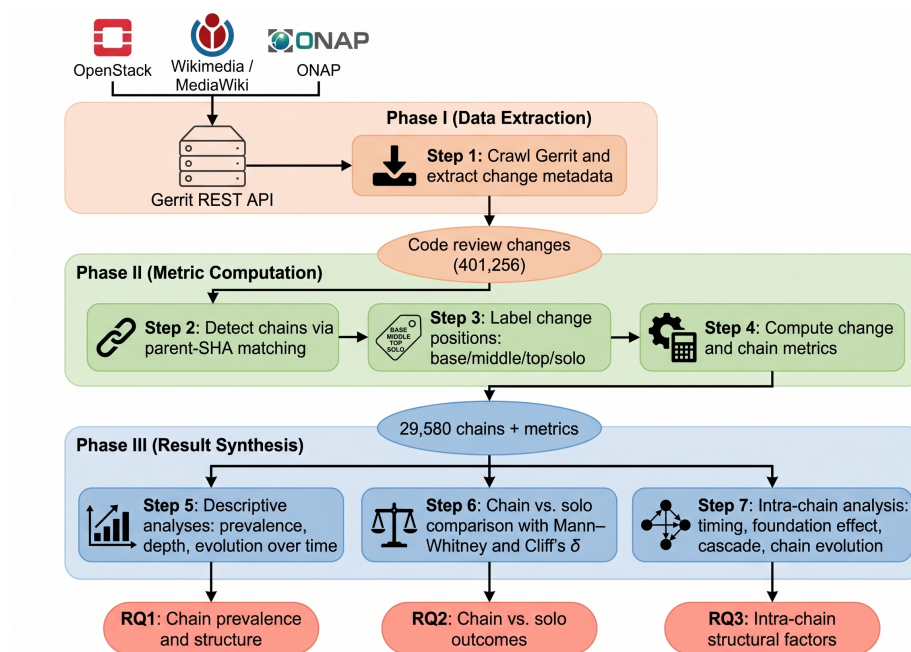
4.1 Research Questions

We organize our study around three research questions that progressively examine relation chains as code-review artifacts: how developers adopt them across projects and over time, how chain membership relates to review outcomes compared to isolated changes, and which intra-chain structural factors shape those outcomes.

RQ1. *[Chain Usage]* What characterizes the usage of relation chains across projects and ecosystems?

RQ2. *[Chain vs. Solo Outcomes]* How do relation chains reshape review coordination, validation workload, and review latency compared to isolated changes?

RQ3. *[Intra-chain Structural Factors]* What structural factors within a chain shape its review outcomes?



■ **Figure 3** Study methodology overview.

RQ1 establishes how often, how deeply, and over what time horizon developers use chains. The basic shape of chain usage (prevalence, depth, temporal stability) determines whether chain-level analysis is broadly relevant or confined to a narrow slice within a few projects. RQ2 examines whether reviews that belong to a relation chain differ systematically from isolated reviews in terms of review duration, rebasing activity, discussion, and CI workload. We compare chain and solo changes within the same size categories and inspect whether similar patterns appear across projects. If chain-associated reviews consistently differ from isolated reviews, then chain-level signals may be useful for future review analytics and tooling. Because the study is observational, we interpret the results as recurring associations rather than direct causal effects of chaining. RQ3 then examines which *intra-chain* features drive the differences identified in RQ2 (chain position, base-change activity, structural evolution), turning the chain from a binary attribute into a set of fine-grained signals that downstream review-support tools (recommendation systems, dashboards, CI schedulers) could consume. Limitations are discussed in Section 7.

4.2 Phase I: Data Extraction

We selected three Gerrit ecosystems (OpenStack, Wikimedia, ONAP) as representative large ecosystems where coordination across dependent changes is needed, and where prior MCR research has established the platform as a reliable empirical setting [29, 23, 39, 8]. The three differ on dimensions likely to interact with chain behaviour: review culture, language and domain, and CI-gating strictness, letting us separate effects that are general properties of chains from those tied to a specific development context. All three use the same review platform with the same chain-detection mechanism (patchset-level parent-SHA links), so cross-ecosystem comparisons rest on a uniform data model. Within each ecosystem, we required publicly accessible review history, at least 1,000 completed changes, and active

development spanning multiple years, following prior MCR conventions [29, 23, 3]. Table 1 summarizes the corpus. For each project, we queried the Gerrit REST API in five passes covering change metadata, patchset revisions with parent SHAs, file-level diffs, review messages (distinguishing humans from CI bots), and review labels.

■ **Table 1** Overview of the studied Gerrit corpus across the OpenStack, Wikimedia, and ONAP ecosystems. Per-project counts of total changes, merged and abandoned outcomes, and reconstructed relation chains.

Ecosystem	Project	Changes	Merged	Aband.	Chains
OpenStack	nova	42,781	31,145	11,636	3,682
	neutron	29,151	22,430	6,603	2,015
	cinder	18,797	13,793	4,094	1,104
	horizon	13,654	10,503	3,058	406
	heat	12,446	10,047	2,299	498
Wikimedia	mw/core	84,626	72,884	10,276	10,036
	ext/Wikibase	28,839	25,692	2,972	3,662
	ext/VE	17,194	16,080	986	1,291
	ext/MF	15,365	13,830	1,501	1,341
	ops/puppet [†]	118,257	110,366	7,891	3,012
ONAP	so	5,208	4,439	699	680
	sdc	4,728	4,157	543	846
	oom	5,743	4,812	904	588
	cps	2,524	2,264	243	184
	ccsdk/cds	1,943	1,809	131	235
Total		401,256			29,580

[†]ops/puppet’s chain count includes one auto-generated chain of depth 59,945 from its config-as-code workflow; we exclude it from aggregate statistics.

4.3 Phase II: Chain Detection and Metric Computation

A *relation chain* is a set of Gerrit changes linked by parent–child dependencies at the *patchset* level. Every patchset is a full git commit with a parent SHA; Gerrit detects a chain link whenever a patchset’s parent SHA matches the latest revision SHA of another in-review change in the same project. Because each change has its own evolving patchset sequence, the parent SHA can change across patchsets, the mechanism by which chains restructure during review (Section 5.3). For chain detection (step 2), we follow Gerrit’s convention and link changes by the parent SHA of each change’s latest patchset. A change with no in-review parent and no in-review children is classified as *solo*.

We label each chain member by position: the *base* has no in-review parent, the *top* has no in-review descendant, and *middle* members are neither. In a chain of $n \geq 3$, exactly one base, one top, and $n - 2$ middle members exist; branching chains produce one top per branch. *Solo* is a fourth label for changes outside any chain. For each change we compute six *per-change* metrics (step 3). **#Revisions** is the number of distinct patchset revisions, from the Gerrit `revisions` endpoint [29, 23, 38]. **#Rebases** is the subset of patchsets whose `kind` is `NO_CHANGE` or `TRIVIAL_REBASE`, isolating dependency-driven rebases from substantive revisions. **Review duration (hours)** is the time from first patchset upload to final state, computed across merged and abandoned changes to avoid survivorship bias [8, 47]. **Discussion messages** is the count of top-level human review messages, excluding accounts on a per-project bot allowlist (top-50 commenters, CI usernames such as `zuul`, `jenkins`) [29, 46]. **Inline comments** is the count of human-authored code-anchored comments, filtered with the same allowlist. **CI/CD jobs** is the count of verification messages from accounts in the bot allowlist [27, 9]. To account for change size as a confound, changes are bucketed by Gerrit size category (XS: ≤ 9 lines; S: 10–29; M: 30–99; L: 100–999; XL: ≥ 1000) [29, 23].

At the chain level (step 4) we compute four *per-chain* metrics. *Chain depth* is the number of members (≥ 2). *Submission gap* and *merge gap* are computed for each consecutive pair (c_i, c_{i+1}) as the time between uploads and between merges. For each chain with at least two merged members, we compute the *base-change* and *descendant-mean* values of each per-change metric, supporting correlation analyses. For chain-evolution analysis we compare each member's parent SHA at Patchset 1 against its parent SHA at the final patchset, classifying any non-trivial-rebase change as a structural evolution event.

4.4 Phase III: Analysis

We use non-parametric tests throughout because code-review metrics are right-skewed [4, 5]. All tests are two-sided. For RQ1's temporal trend (step 5), monthly prevalence violates independence because consecutive months share contributors, release cadences, and backlogs. We therefore use the Mann–Kendall test [14], robust to autocorrelation, and report Sen's slope [37] in percentage points per year (pp/yr), both established in SE temporal studies [42, 22]. H_0 : no monotonic trend; H_1 : an increasing or decreasing trend.

For chain-vs-solo comparisons (RQ2, step 6), we report Mann–Whitney U [28, 14] and Cliff's δ [13], using Romano et al.'s thresholds [36]: $|\delta| < 0.147$ negligible, < 0.33 small, < 0.474 medium, and ≥ 0.474 large. H_0 : a random chain member and solo change are equally likely to be larger; H_1 : one is stochastically greater. Because large samples can make significance misleading, we treat Cliff's δ as the primary measure of practical magnitude and emphasize directional consistency across projects and ecosystems.

For per-chain base-vs-descendant correlations (RQ3, step 7), we report Spearman's ρ with two-sided p -values [14], which captures monotonic association without assuming linearity or normality. H_0 : base and descendant-mean metrics are rank-independent; H_1 : they are monotonically associated.

Significance markers are *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and ns otherwise. We apply no multiple-comparison correction because each project is an independent observational unit and interpret each effect by its sign and magnitude.

5 Results

5.1 RQ1: Prevalence and Structure

Table 2 summarizes per-project chain-prevalence trends; Figure 4 illustrates two contrasting exemplars. Table 3 reports per-project chain-depth statistics, and Figure 5 shows monthly mean chain depth per project.

Finding #1.1: Relation-chain prevalence varies across projects and ecosystems. Excluding the `operations/puppet` outlier, the proportion of changes in a chain ranges from 5.1% in OpenStack `Horizon` to 48.5% in ONAP `sdc`, with a median of 17% across the remaining 14 projects. Variation differs by ecosystem: ONAP 11–49%, OpenStack 5–32%, Wikimedia 12–22%. Chain usage is therefore not uniform but reflects differences in development practices and tooling conventions.

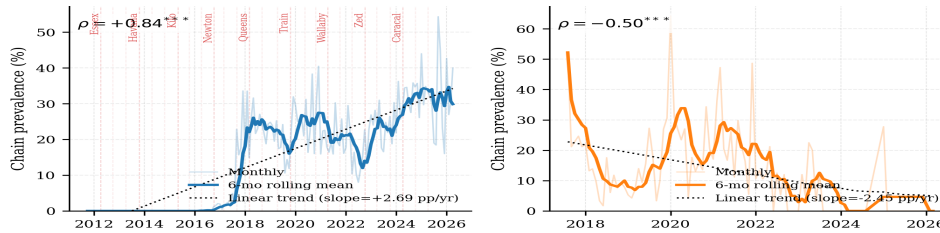
Finding #1.2: Relation-chain adoption evolves over time, with most projects showing increasing usage. The Mann–Kendall trend test (Table 2) yields significant monotonic trends in 14 of 15 projects: 10 increasing, 4 decreasing. The only project without a significant trend is `ccsdk/cds`, which has the shortest observation window (38 months). All OpenStack

■ **Table 2** Per-project chain-prevalence trends (Mann–Kendall test on monthly prevalence series). *Chained %* is the proportion of project changes that belong to a relation chain. Sen’s slope reports the trend magnitude in percentage points per year.

Ecosystem	Project	Chained %	Months	Sen’s slope (pp/yr)	MK <i>p</i>
OpenStack	Nova	31.5%	175	+1.83	***
	Neutron	10.8%	176	+2.44	***
	Cinder	11.0%	168	+2.81	***
	Horizon	5.1%	174	+0.76	***
	Heat	7.5%	159	+2.23	***
Wikimedia	mw/core	19.1%	170	+1.93	***
	ext/Wikibase	21.9%	167	+2.76	***
	ext/VisualEditor	13.2%	170	+1.54	***
	ext/MobileFrontend	12.1%	170	+0.43	***
	ops/puppet [†]	97.3%	170	+0.27	***
ONAP	so	27.1%	81	−3.42	**
	sdc	48.5%	89	−8.02	***
	oom	16.9%	104	−1.99	***
	cps	11.2%	65	−1.53	*
	ccsdk/cds	25.0%	52	−1.59	ns

[†]ops/puppet’s near-100% prevalence reflects auto-generated config-as-code chains, not human-authored review.

projects exhibit increasing trends (+0.8 to +2.8 pp/yr, $p < 0.001$), and four of five Wikimedia projects follow the same pattern. Several ONAP projects display decreasing trends (−8.0 to −1.5 pp/yr), notably *oom* (Figure 4b), where monthly prevalence drops from over 60% in 2018–2019 to below 20% by 2024.



■ **Figure 4** Monthly chain-prevalence exemplars: (a) increasing trend in OpenStack Neutron; (b) decreasing trend in ONAP *oom*. Bold line: 6-month rolling mean; dotted line: linear trend.

ONAP *oom* change 137529,³ which introduces Kafka support across multiple policy charts in a single 2024 change, illustrates this evolution. Similar cross-component modifications in earlier years were often decomposed into linked chains, suggesting a gradual shift away from chain-oriented coordination; we revisit potential explanations in Section 6.

Finding #1.3: Relation chains are dominated by short pair-like structures, with project-specific deep tails and one auto-generated outlier. Per-project depth statistics show similar central tendencies across ecosystems (Table 3): 12 of 15 projects have median depth 2, 14 of 15 have Q1 of 2, and 11 of 15 have $Q3 \leq 4$. Thus, most chains are pair-like structures or short sequences of two to four members. Mean depth is low (2.3–6.1), with standard deviation below 4 in 12 of 15 projects. The tails vary substantially: maximum depth ranges from 5 (*cps*) to 98 (*sdc*), and seven projects contain at least one chain deeper than 30. ONAP *sdc* has the deepest profile (mean 6.1, Q3 of 7, maximum 98), whereas *cps* has

³ <https://gerrit.onap.org/r/c/oom/+137529>

maximum depth 5 and standard deviation 0.6. The `ops/puppet` project is an outlier: its maximum depth of 59,945 comes from a single auto-generated chain in its config-as-code workflow (Table 1), which we exclude from aggregates.

■ **Table 3** Per-project chain-depth statistics. `ops/puppet`’s maximum reflects a single auto-generated config-as-code chain (Table 1).

Ecosystem	Project	<i>n</i>	Min	Q1	Median	Mean	Q3	Max	Std
OpenStack	nova	4,985	2	2	2	2.7	3	56	1.8
	neutron	2,015	2	2	2	3.1	3	26	2.5
	cinder	1,104	2	2	2	4.8	5	38	5.4
	horizon	406	2	2	2	3.6	4	22	3.3
	heat	498	2	2	2	3.5	4	29	3.0
Wikimedia	mw/core	10,036	2	2	2	3.2	4	31	2.1
	ext/Wikibase	3,662	2	2	2	3.4	4	34	2.8
	ext/VisualEditor	1,291	2	2	2	3.0	3	12	1.6
	ext/MobileFrontend	1,341	2	2	3	3.4	4	22	2.3
	ops/puppet	3,012	2	3	5	38.2	9	59,945	1244.4
ONAP	so	680	2	2	2	3.2	4	21	2.3
	sdc	846	2	2	3	6.1	7	98	7.8
	oom	588	2	2	2	3.4	3	46	3.6
	cps	184	2	2	2	2.3	2	5	0.6
	ccsdk/cds	235	2	2	2	3.5	4	18	2.7

Finding #1.4: Mean chain depth shows no systematic temporal trend in most projects.

Grouping chains by the month their base was created (Figure 5), Spearman correlations between calendar month and monthly mean depth range from -0.34 to $+0.33$ and are non-significant in 12 of 15 projects. Exceptions are ONAP `so` ($\rho = -0.57$, $p < 0.001$) and `sdc` ($\rho = -0.53$, $p < 0.001$), where chains become shallower over time, consistent with declining prevalence in Finding #1.2, and OpenStack Cinder, which shows the opposite trend ($\rho = +0.32$, $p < 0.01$). Thus, outside these cases, chain *shape* remains stable while *adoption* grows: the increase in Finding #1.2 reflects more chains, not longer ones.

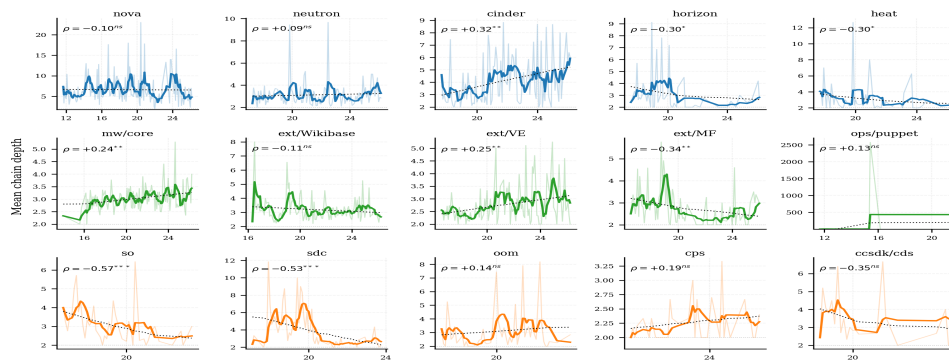
Finding #1.5: Depth spikes correspond to coordinated engineering campaigns, not noise.

Figure 5 also reveals months where mean depth rises to several times the project’s long-run median. We identified 14 spike months (mean depth above the project median plus two standard deviations); 13 are multi-chain bursts in which three to nineteen chains, several deep, are created together. For example, OpenStack Horizon created seventeen chains in December 2018 with a mean depth of nine and a maximum of twenty-two, driven by a single author decomposing a project-wide `pylint` cleanup⁴ into dependent chains. Similar spikes occur during architectural migrations and refactoring sweeps (full list in the replication package), indicating that depth spikes mark coordinated multi-chain campaigns.

RQ1 summary

Relation chains are a common but ecosystem coordination mechanism in Gerrit-based development, with prevalence ranging from 5% to 49% across projects (median 17%). Chain adoption is increasing in most projects. Chains are typically shallow (median depth 2), and their structural characteristics remain stable despite changing adoption rates. Deeper chains, when they occur, are associated with engineering workflows such as sequential refactoring and diagnostic branching.

⁴ <https://review.opendev.org/q/topic:pylint+project:openstack/horizon>



■ **Figure 5** Monthly mean chain depth per project (6-month rolling mean; dotted line = linear trend). Horizontal dotted line at depth 2 marks the pair floor.

5.2 RQ2: Review Outcomes

Table 5 reports descriptive statistics and Mann–Whitney effect sizes for chain-vs-solo comparisons across six metrics and five Gerrit size buckets. Table 4 breaks down the same metrics by chain-position label. Figure 6 shows per-project patchset count by chain position.

Finding #2.1: Relation chains introduce synchronization overhead through additional revisions, rebasing activity, and longer review duration. Across all five Gerrit size buckets, chain members take a median of $2.6\times$ longer to merge and undergo $1.5\times$ more revision rounds than comparable solo changes (Mann–Whitney U, $p < 0.001$; Table 5). Cliff’s δ is *small* in four of five buckets and *large* only in the XL bucket ($\delta = 0.49$), showing that the strongest overhead is concentrated among very large changes. This reflects repeated rebasing, validation, and merge-order coordination across dependencies. The OpenStack Horizon December 2018 `pylint` cleanup⁵, which created seventeen dependent chains in one month, illustrates this effect: each chain accumulated its own review iterations plus dependency-related coordination overhead.

Finding #2.2: Chain members do not attract more top-level discussion than solo changes, but receive significantly more code-anchored inline comments in OpenStack and Wikimedia. In OpenStack and Wikimedia, solo changes attract more median discussion messages than middle- or top-of-chain members (Table 4). In contrast, OpenStack chain members receive significantly more inline comments across all five size buckets (Cliff’s $\delta = 0.21\text{--}0.34$, all *small*, $p < 0.001$; Table 5). Wikimedia shows significant differences in every bucket but with mixed direction, while ONAP shows the smallest and least consistent effect. Overall, feedback appears to shift from top-level discussion toward code-anchored scrutiny and synchronization/validation maintenance.

Finding #2.3: Relation chains exhibit a CI amplification effect under strict gating regimes. Chain members run a median of 10–23 CI/CD jobs depending on position, against fewer than two for solo changes in the same ecosystem (Table 4). The split is largest in OpenStack, where base, middle, and top members each trigger a test pipeline on every patchset under the project’s gating policy [27, 9]; the accumulation matches the multi-hour merge gaps

⁵ <https://review.opendev.org/q/topic:pylint+project:openstack/horizon>

reported in Finding #3.1. ONAP, with lighter CI gating, shows the smallest split (chain CI medians of 0–2 versus solo medians of 2). The OpenStack Neutron [OVN] migration of November 2019⁶, with nine dependent chains advancing the same migration, illustrates the cumulative cost: each member’s revisions retrIGGERED the CI suite. We call this recurring downstream validation workload the *CI amplification effect*, where revisions to one chain member trigger repeated validation across dependent descendants.

■ **Table 4** Per-ecosystem chain-position breakdown of six review metrics. For each ecosystem (OpenStack, Wikimedia, ONAP), each metric is split across the three chain-position labels (Base, Middle, Top) and summarised as Min, Q1, Median, Mean, Q3, Max. `ops/puppet` is excluded from the Wikimedia rows (see Table 1).

	Metric	Base						Middle						Top					
		Min	Q1	Med	Mean	Q3	Max	Min	Q1	Med	Mean	Q3	Max	Min	Q1	Med	Mean	Q3	Max
OpenStack	#Revisions	1	1	3	5.61	6	126	1	2	3	5.72	7	103	1	1	3	4.78	5	114
	#Rebases	0	0	2	4.61	5	125	0	1	2	4.72	6	102	0	0	2	3.78	4	113
	Review dur. (h)	0.18	63.9	270	1167	990	60,844	0.28	116	362	1181	1143	47,431	0.40	105	324	1011	964	38,658
	Discussion msg.	0	1	8	30.4	27	1,796	0	0	4	25.4	19	951	0	0	4	15.6	16	1,185
	Inline comments	0	7	17	38.6	40	1,786	0	9	19	36.1	39	935	0	7	15	29.1	33	1,185
	CI/CD jobs	0	5	10	21.5	23.2	550	0	6	14	23.0	28	408	0	5	11	19.2	23	581
Wikimedia	#Revisions	1	1	3	5.04	6	180	1	2	3	4.85	6	121	1	1	2	3.56	4	198
	#Rebases	0	0	2	4.04	5	179	0	1	2	3.85	5	120	0	0	1	2.56	3	197
	Review dur. (h)	0	2.95	33.5	489	176	85,686	0.00	4.75	43.9	333	155	43,391	0	4.56	31.8	366	147	47,118
	Discussion msg.	2	10	16	26.1	29	765	2	10	16	23.4	27	505	1	9	13	17.8	21	740
	Inline comments	0	2	6	10.1	11	405	0	2	5	9.61	12	312	0	1	4	7.50	10	380
	CI/CD jobs	0	3	9	16.0	18	547	0	3	9	13.8	17	353	0	3	7	10.3	13	530
ONAP	#Revisions	1	1	2	3.59	4	77	1	2	2	2.95	3	50	1	1	2	2.78	3	41
	#Rebases	0	0	1	2.59	3	76	0	1	1	1.95	2	49	0	0	1	1.78	2	40
	Review dur. (h)	0.00	2.85	24.4	203	146	6,353	0.01	3.52	29.6	120	117	4,625	0.02	3.66	23.3	122	116	2,562
	Discussion msg.	2	8	13	30.3	29	820	2	8	11	16.0	16	567	2	7	11	20.4	19	655
	Inline comments	0	7	11	25.0	23	748	1	8	9	14.1	14	511	0	6	9	17.2	16	604
	CI/CD jobs	0	0	2	5.22	6	347	0	0	0	1.92	3	96	0	0	2	3.16	4	107

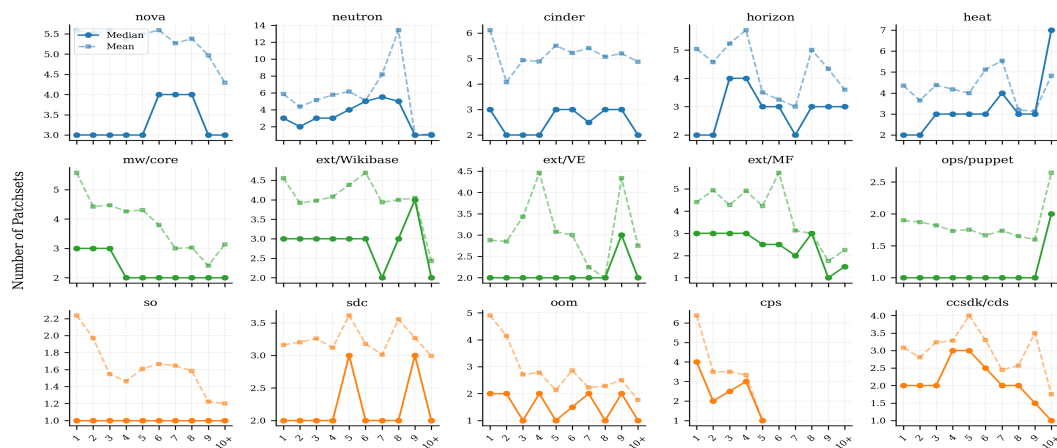
Finding #2.4: Middle-of-chain members function as synchronization bottlenecks within dependency-linked review workflows. Middle-of-chain members function as synchronization bottlenecks within dependency-linked review workflows. Across all three ecosystems, middle-of-chain changes show higher median #Revisions, more #Rebases, and longer review duration than the base or the top of the same chain (Table 4). The per-project view confirms the pattern: in nearly every project, #Revisions medians at middle positions sit at or above those at the base and the top (Figure 6). This matches bottom-up merge enforcement: a middle member is simultaneously blocked by its predecessor and blocking its successor, accumulating the iteration cost of waiting for changes below while continuing to receive feedback from above. As a result, middle members effectively function as synchronization bottlenecks that absorb the highest coordination burden within the chain. Reviewers internalise this dependency: on a 17-member ONAP `sdc` chain (change 88612⁷), one held back approval with “*waiting for parent to be verified.*”

⁶ <https://review.opendev.org/q/topic:ovn-migration+project:openstack/neutron>

⁷ <https://gerrit.onap.org/r/c/sdc/+88612>

Table 5 Per-ecosystem descriptive statistics and effect sizes for review metrics, comparing chain (Ch) vs. solo (So) changes across Gerrit size buckets. p -values from Mann–Whitney U test; Cliff's δ magnitude: N = negligible, S = small, M = medium, L = large.

		Min		Q1		Mean		Median		Q3		Max					
Ecosystem	Metric	Size	Ch	So	Ch	So	Ch	So	Ch	So	Ch	So	Ch	So	p	Cliff's δ	
OpenStack	#Revisions	XS	1	1	1	2.349	1.889	1	1	3	2	56	229	<0.001	0.13	(N)	
		S	1	1	1	3.662	2.753	2	2	4	3	63	89	<0.001	0.17	(S)	
		M	1	1	1	4.772	4.011	3	2	6	5	91	106	<0.001	0.10	(N)	
		L	1	1	2	2	8.200	7.481	5	4	10	9	126	155	<0.001	0.08	(N)
		XL	1	1	3	1	10.092	8.222	6	3	12	10	105	176	<0.001	0.21	(S)
	#Rebases	XS	0	0	0	0	0.736	0.266	0	0	1	0	45	175	<0.001	0.20	(S)
		S	0	0	0	0	1.318	0.428	0	0	1	0	51	67	<0.001	0.27	(S)
		M	0	0	0	0	1.670	0.698	0	0	2	0	59	75	<0.001	0.26	(S)
		L	0	0	0	0	2.606	1.432	1	0	3	1	65	66	<0.001	0.23	(S)
		XL	0	0	0	0	2.006	0.901	1	0	2.250	0	36	54	<0.001	0.33	(S)
	Review dur. (hours)	XS	0.2	0	24.1	13.4	486.179	334.667	106.167	60.036	365.022	219.980	23673.56582.8	<0.001	0.15	(S)	
		S	0.2	0	67.26	2	798.442	539.740	213.154	116.154	672.892	401.990	32944.7	62328.2	<0.001	0.20	(S)
		M	0.2	0	99.4	45.4	1067.866	737.522	323.892	180.072	982.629	654.940	41921.9	59373.7	<0.001	0.18	(S)
		L	0.2	0	180.2	69.2	1646.194	1077.740	566.216	318.616	1675.422	1069.743	60844.4	58980.3	<0.001	0.21	(S)
		XL	0.3	0	202.3	10.7	2081.738	643.074	755.262	95.516	2108.616	524.222	47430.6	23719.5	<0.001	0.48	(L)
	Discussion messages	XS	0	0	5	6	15.485	19.556	10	11	19	23	283	3966	<0.001	-0.12	(N)
		S	0	0	7	8	23.843	27.665	14	16	29	32	368	1777	<0.001	-0.05	(N)
		M	0	0	8	9	30.613	39.130	17	20	36	44	748	3372	<0.001	-0.10	(N)
		L	0	0	11	11	51.410	70.236	25	29	59	75	1348	3314	<0.001	-0.07	(N)
		XL	0	0	13	8	86.777	84.590	33	24	84	79	1786	3525	<0.001	0.09	(N)
	Inline comments	XS	0	0	0	2	12.529	17.906	4	10	15	22	288	3966	<0.001	-0.22	(S)
		S	0	0	0	4	14.514	24.324	3	12	15	29	580	1777	<0.001	-0.31	(S)
		M	0	0	0	5	17.201	33.623	4	15	16	37	752	3372	<0.001	-0.34	(M)
		L	0	0	2	8	37.134	62.364	10	23	34	62	1360	3314	<0.001	-0.25	(S)
		XL	0	0	2	7	89.455	82.035	20	21	98.500	76.750	1796	3525	0.230	-0.03	(N)
	CI/CD jobs	XS	0	0	3	0	10.463	4.295	7	0	13	6	281	848	<0.001	0.44	(M)
		S	0	0	5	0	15.741	6.063	10	0	20	8	332	276	<0.001	0.50	(L)
		M	0	0	6	0	19.870	8.637	12	0	24	10	387	1344	<0.001	0.49	(L)
		L	0	0	6	0	30.304	12.266	16	0	37	10	581	592	<0.001	0.52	(L)
		XL	0	0	5	0	33.041	7.712	14	0	38.250	2	444	621	<0.001	0.64	(L)
Wikimedia	#Revisions	XS	1	1	1	1	1.977	1.570	2	1	2	2	59	56	<0.001	0.25	(S)
		S	1	1	1	1	2.797	2.188	2	1	3	3	53	81	<0.001	0.25	(S)
		M	1	1	2	1	3.972	2.861	3	2	5	4	102	76	<0.001	0.27	(S)
		L	1	1	2	1	6.587	3.983	4	2	8	5	198	152	<0.001	0.35	(M)
		XL	1	1	2	1	9.220	3.341	4	1	10	2	180	229	<0.001	0.55	(L)
	#Rebases	XS	0	0	0	0	0.632	0.161	0	0	1	0	36	29	<0.001	0.37	(M)
		S	0	0	0	0	0.833	0.274	1	0	1	0	34	36	<0.001	0.40	(M)
		M	0	0	0	0	1.031	0.371	1	0	1	0	40	57	<0.001	0.38	(M)
		L	0	0	0	0	1.343	0.484	1	0	2	0	45	71	<0.001	0.39	(M)
		XL	0	0	0	0	1.388	0.338	1	0	2	0	30	50	<0.001	0.44	(M)
	Review dur. (hours)	XS	0	0	0	0.2	89.256	139.776	0.193	1.576	13.016	20.944	47162.1	79040.4	<0.001	-0.25	(S)
		S	0	0	0	0.4	164.258	235.849	2.222	5.796	50.216	67.619	53249.6	52312.5	<0.001	-0.11	(N)
		M	0	0	0.6	0.3	291.977	299.865	19.830	11.239	119.545	99.448	59875.1	93770.5	<0.001	0.07	(N)
		L	0	0	4.5	0.1	452.677	294.488	66.417	4.396	241.956	110.864	85686.3	77280.7	<0.001	0.29	(S)
		XL	0	0	2.6	0.2	693.852	145.923	59.204	0.386	311.383	6.074	40490.2	29884.2	<0.001	0.51	(L)
	Discussion messages	XS	0	0	2	2	4.194	5.868	3	6	5	8	185	252	<0.001	-0.25	(S)
		S	0	0	2	2	5.686	7.119	4	6	7	9	143	231	<0.001	-0.12	(N)
		M	0	0	2	2	7.576	8.454	5	6	10	11	159	171	<0.001	-0.03	(N)
		L	0	0	3	2	11.941	10.535	7	5	14	12	274	395	<0.001	0.12	(N)
		XL	0	0	3	0	17.220	7.265	7	1	17	5.500	405	681	<0.001	0.48	(L)
	Inline comments	XS	1	1	4	6	7.109	8.834	6	7	8	10	296	252	<0.001	-0.27	(S)
		S	1	1	5	6	10.309	11.319	8	9	12	14	200	317	<0.001	-0.10	(N)
		M	1	1	7	6	15.200	13.778	11	10	18	17	400	303	<0.001	0.08	(N)
		L	1	1	9	6	26.174	17.419	16	9	32	21	740	443	<0.001	0.28	(S)
		XL	2	1	8	4	38.059	13.606	16	6	39	10	765	832	<0.001	0.48	(L)
	CI/CD jobs	XS	0	0	0	0	2.915	2.965	3	0	4	5	282	173	<0.001	0.14	(N)
		S	0	0	0	0	4.623	4.200	3	0	6	6	186	211	<0.001	0.13	(N)
		M	0	0	0	0	7.625	5.324	4	0	10	7	309	261	<0.001	0.21	(S)
		L	0	0	0	0	14.234	6.884	7	0	17	6	530	372	<0.001	0.30	(S)
		XL	0	0	0	0	20.839	6.341	7	4	18	5	547	675	<0.001	0.25	(S)
ONAP	#Revisions	XS	1	1	1	1	1.911	1.566	1	1	2	2	19	22	<0.001	0.16	(S)
		S	1	1	1	1	2.432	2.025	2	1	3	2	27	47	<0.001	0.19	(S)
		M	1	1	1	1	2.850	2.561	2	2	3	3	23	43	<0.001	0.14	(N)
		L	1	1	1	1	4.011	4.077	2	2	4	5	77	103	0.018	0.04	(N)
		XL	1	1	2	1	5.099	4.343	3	2	6	5	42	72	<0.001	0.20	(S)
	#Rebases	XS	0	0	0	0	0.567	0.181	0	0	1	0	10	17	<0.001	0.25	(S)
		S	0	0	0	0	0.764	0.248	1	0	1	0	10	10	<0.001	0.33	(M)
		M	0	0	0	0	0.802	0.327	1	0	1	0	12	12	<0.001	0.34	(M)
		L	0	0	0	0	0.953	0.457	1	0	1	1	18	13	<0.001	0.31	(S)
		XL	0	0	0	0	1.081	0.503	1	0	1	0	12	23	<0.001	0.39	(M)
	Review dur. (hours)	XS	0	0	1.9	2.1	107.354	91.794	15.056	18.035	71.062	71.606	4306.5	14296	0.331	-0.02	(N)
		S	0	0	3	3.3	137.344	109.260	20.265	21.815	91.860	91.851	6031.2	14849	0.838	-0.00	(N)
		M	0	0	3.3	4.1	125.871	136.358	25.111	24.938	118.099	117.314	4698.9	10559.6	0.603	-0.01	(N)
		L	0	0	5.1	10.6	174.648	223.469	43.306	51.805	170.484	192.646	6353.3	10430.8	<0.001	-0.06	(N)
		XL	0	0	7	4.7	267.763	191.963	70.475	38.814	261.642	167.277	5582.2	5708.4	<0.001	0.12	(N)
	Discussion messages	XS	0	0	6	6	12.801	12.855	8	9	13	14	181	227	0.256	-0.02	(N)
		S	0	0	7	7	15.255	16.322	9	9	15	17	235	405	0.388	-0.02	(N)
		M	1	0	7	7	16.821	18.394	10	10	16	20	268	430	0.460	-0.02	(N)
		L	0	0	7	7	24.363	27.364	11	13	21	31	748	530	<0.001	-0.07	(N)
		XL	0	0	9	7	25.774	26.381	14	13	26	27	473	626	0.016	0.08	(N)
	Inline comments	XS	2	1	7	7	14.719	15.893	9	11	14	18	204	259	<0.001	-0.11	(N)
		S	2	2	8	8	17.544	20.147	10	12	17	22	250	454	<0.001	-0.10	(N)
		M	2	2	8	8	19.892	23.796	11	13	18	27	305	476	<0.001	-0.10	(N)
		L	2	2	8	9	29.261	35.821	13	17	26	41	820	968	<0.001	-0.12	(N)
		XL	2	2	10	9	30.737	33.516	16	15	31	33	518	712	0.287	0.03	(N)
	CI/CD jobs	XS	0	0	0	0	1.918	3.038	2	2	3	3	39	64	<0.001	-0.22	(S)
		S	0	0	0	1	2.289	3.824	2	2	3	4	44	74	<0.001	-0.24	(S)
		M	0	0	0	2	3.071	5.402	2	3	4	6	107	108	<0.001	-0.26	(S)
		L	0	0	0	2	4.897	8.457	2	3	5	9	347	543	<0.001	-0.22	(S)
		XL	0	0	0	2	4.964	7.135	1	2	6	7	186	164	<0.001	-0.14	(N)



■ **Figure 6** Patchset count by chain position-from-base, per project. Solid line: median; dashed line: mean.

RQ2 summary

Relation chains introduce synchronization overhead through increased revisions, rebasing activity, review duration, and amplified CI workload compared to isolated changes (Finding #2.1). These coordination costs emerge primarily from dependency management rather than increased conversational review activity (Finding #2.2). Under strict CI-gating regimes, relation chains exhibit a *CI amplification effect*, where revisions to one member repeatedly retrigger validation across dependent descendants (Finding #2.3). Within chains, middle members function as synchronization bottlenecks that absorb the highest coordination burden due to simultaneous upstream and downstream dependency constraints (Finding #2.4).

5.3 RQ3: Structural Factors

Table 6 reports submission and merge gaps between consecutive chain members. Figure 7 shows Spearman correlations between base-change and descendant-mean review metrics.

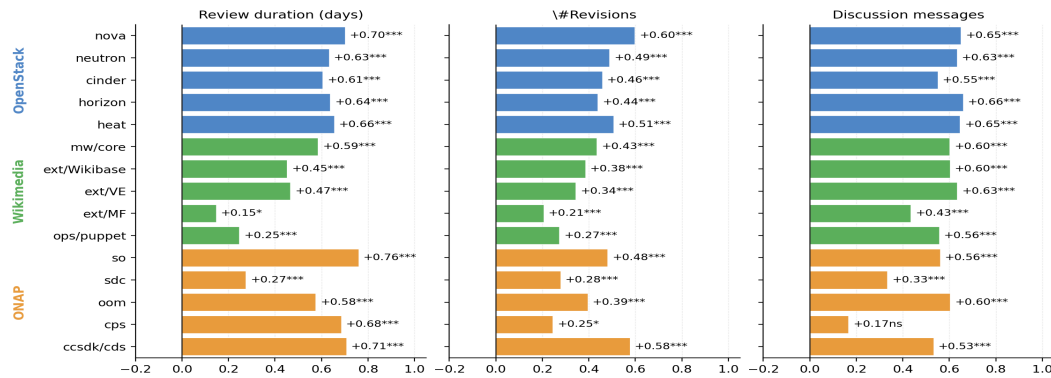
Finding #3.1: Chain throughput is set by the project's CI gating regime rather than chain mechanics. Relation chains unfold in two timing regimes: a *submission gap* (time between uploads of consecutive members) and a *merge gap* (time between merges). Submission is fast on aggregate: pooled median 14 minutes, with project medians from near-zero in OpenStack Horizon and Heat to roughly six hours in ONAP sdc and cps (Table 6); these reflect developer work patterns, since a multi-step refactor prepared offline appears within minutes while incremental uploads widen the spread. Merge gaps separate ecosystems more sharply: medians sit below one hour for most Wikimedia and ONAP projects but between two and ten hours for every OpenStack project (Horizon 2.23 h to Neutron 9.82 h). The split tracks CI-gating regimes [27, 9]: OpenStack runs a per-member test pipeline before each descendant's submission completes, while Wikimedia and ONAP permit descendants to follow within minutes once the base has merged. The Horizon pylint cleanup of December 2018⁸ shows this at chain level: chains assembled offline (small submission gaps) produced merge gaps in hours because each descendant waited for its predecessor's CI pipeline on the rebased revision. The same dependency structure therefore produces substantially different coordination costs under different CI-gating regimes.

⁸ <https://review.opendev.org/q/topic:pylint+project:openstack/horizon>

■ **Table 6** Per-project descriptive statistics of submission and merge gaps (in hours) between consecutive chain members. `ops/puppet` is excluded from aggregate statistics (see Table 1).

Ecosystem	Project	Submission gap (h)						Merge gap (h)					
		Min	Q1	Med	Mean	Q3	Max	Min	Q1	Med	Mean	Q3	Max
OpenStack	nova	0.00	0.00	0.10	222.7	21.06	41,640	0.00	0.02	4.62	175.2	65.80	17,956
	neutron	0.00	0.00	0.56	267.6	74.75	22,172	0.00	0.06	9.82	124.1	72.69	24,983
	cinder	0.00	0.00	0.36	525.9	96.41	62,353	0.00	0.01	4.35	149.8	69.50	7,079
	horizon	0.00	0.00	0.00	455.6	35.78	14,003	0.00	0.04	2.23	129.5	70.95	2,520
	heat	0.00	0.00	0.00	268.6	18.12	9,691	0.00	0.06	4.81	263.2	72.62	19,390
Wikimedia	mw/core	0.00	0.00	0.11	229.7	4.02	49,514	0.00	0.00	0.09	23.3	1.17	8,268
	ext/Wikibase	0.00	0.00	0.80	104.5	21.32	38,206	0.00	0.01	0.27	22.5	2.91	7,255
	ext/VE	0.00	0.00	0.25	184.8	6.46	45,289	0.00	0.00	0.06	36.8	1.96	8,353
	ext/MF	0.00	0.00	0.31	72.3	21.07	4,635	0.00	0.03	0.53	18.9	8.11	601
	ops/puppet [†]	0.00	0.26	1.32	155.0	24.66	53,250	0.00	0.10	0.28	2.74	0.74	3,984
ONAP	so	0.00	0.06	0.31	35.9	2.15	7,323	0.00	0.00	0.005	7.1	0.11	1,064
	sdc	0.00	0.45	5.83	84.9	71.23	3,474	0.00	0.04	0.63	6.7	3.31	555
	oom	0.00	0.00	0.21	61.8	18.26	2,809	0.00	0.00	0.007	31.8	1.05	1,372
	cps	0.00	0.29	5.83	90.9	100.79	959	0.00	0.19	0.65	14.3	2.40	311
	ccsdk/cds	0.00	0.02	0.98	37.6	17.48	1,468	0.00	0.00	0.006	5.4	0.56	242

[†]Excluded from aggregate statistics (see Table 1).



■ **Figure 7** Foundation effect: per-chain Spearman correlation between base-change and descendant-mean review metrics, computed within each project.

Finding #3.2: Higher review activity on base is associated with higher review activity on chain descendants. Spearman correlations between the base change’s review metrics and the mean of its descendants’ metrics are positive across all three metrics: review duration (median $\rho = +0.61$, significant in 15/15 projects), discussion messages ($\rho = +0.60$, 14/15), and #Revisions ($\rho = +0.43$, 15/15) (Figure 7). We refer to this regularity as the *foundation effect*: chains with high-effort bases have high-effort descendants, and chains with low-effort bases have low-effort descendants. Because our metrics measure review activity rather than code quality, we describe the correlation as review-effort propagation, not evidence of base quality. The Nova SDK migration chain (change 659691)⁹ illustrates the pattern: its base accumulated 38 #Revisions and 17 discussion messages, and its 14 merged descendants each required additional review activity.

⁹ <https://review.opendev.org/c/openstack/nova/+/659691>

RQ3 summary

The median merge gap between consecutive chain members varies by roughly three orders of magnitude across ecosystems and is explained by the project’s CI-gating regime rather than chain mechanics (Finding #3.1). The *foundation effect*: base-change review activity correlates positively with descendant review activity (Spearman $\rho = 0.43\text{--}0.61$) in 14–15 of 15 projects (Finding #3.2).

6 Discussions and Implications

Although many observed effect sizes are small according to Romano et al.’s Cliff’s δ thresholds, the directional consistency of the results across projects and ecosystems suggests that relation-chain membership captures recurring coordination patterns rather than isolated project-specific anomalies.

Implications for Researchers. The change-level unit of analysis common in code-review research does not capture review dynamics in ecosystems where a substantial share of changes belong to chains; modern review workflows increasingly operate as dependency-aware coordination systems rather than collections of independent review units. When a chain is present, the review process involves decisions across dependent changes (which member to review first, whether to wait for a predecessor, how to handle merge ordering) invisible from any single change. Studies of reviewer load, reviewer assignment [40, 31], and review latency [47, 48] that treat reviews independently may not represent the full process, and predictive models built on change-level features alone omit chain-level signals (chain position, depth, base-change activity, chain dynamism) that our data show correlate with the same outcomes they aim to predict.

Two further chain-level properties deserve attention. First, chains are not static dependency graphs but continuously evolve throughout review, a behaviour we refer to as *dynamic dependency restructuring*: 33.5% of chain members undergo a structural change in their parent SHA before merging, most often by detaching from an in-review parent (40% of evolution events), with rates of 25.0% (OpenStack), 39.7% (Wikimedia), and 50.2% (ONAP). Reviewers describe these dynamics in our metadata: on Wikimedia `mediawiki/core` change 530540¹⁰, one noted that “*the chain is now just blocked on the MediaWikiTestCase-Base thing*,” treating the chain as a unit gated by a single member. Second, chains persist across multi-year timespans: submission gaps reach ~ 4.8 years in OpenStack Nova and merge gaps ~ 2.85 years in Neutron (Table 6). Chain-aware tooling must continuously recompute dependency topology throughout the review lifecycle.

The foundation effect (Finding #3.2) shows base-change review activity correlates with descendant-change activity. Possible mechanisms include scrutiny applied to descendants of a heavily reviewed base, descendants inheriting architectural decisions, and reviewers accumulating context across the chain; our observational data does not distinguish between them, and because our metrics measure review activity rather than code correctness we do not interpret the correlation as evidence of base quality. LLM-based review agents [25, 41, 21] process changes independently of chain dependency, yet chain position is associated with differences in review effort (Finding #2.4) and one third of chain members restructure during review; a natural extension is to provide such tools with chain-level inputs: position, base

¹⁰<https://gerrit.wikimedia.org/r/c/mediawiki/core/+530540>

review state, and recent dependency-graph evolution. Chain prevalence, evolution rates, and median merge gaps also vary across ecosystems by amounts comparable to chain-vs-solo differences within a single ecosystem (Findings #1.2, #3.1), correlating with submit-type and CI-gating configuration [27, 9]; replications should evaluate chain practices per setting rather than pooling across ecosystems.

Implications for Practitioners. Our findings argue against the current default in review tooling, which processes each change as an independent unit [40, 31, 25, 41]: when changes belong to a relation chain, their review outcomes are quantitatively linked. The foundation effect (Finding #3.2) is the central evidence: Spearman $\rho = 0.43\text{--}0.61$ in 14–15 of 15 projects show base review activity predicts descendant activity. The practical consequence is a single high-leverage intervention point per chain. Concentrating reviewer attention on the base (through earlier reviewer assignment, dedicated review slots, or AI-assisted summarisation of the base’s open discussion) is associated with lower per-descendant review effort, multiplying the effect across every member. The Nova SDK migration chain (change 659691) illustrates the scale: 38 #Revisions and 17 discussion messages on the base, with 14 merged descendants each absorbing additional review activity that would likely have been smaller had the base resolved with less friction.

Within chains, middle-of-chain changes show the highest median review effort (Finding #2.4), since a middle change is simultaneously blocked by its predecessor and blocking its successor. Review dashboards and reviewer-assignment systems [40, 31] could use chain position as an input, prioritising changes whose merge unblocks the largest number of downstream changes. Chain-aware CI could reorder test-resource scheduling. Consider the Horizon `pylint` cleanup of December 2018: its seventeen chains were assembled in one afternoon (Finding #3.1), but descendants queued behind each other for hours, each waiting for its predecessor’s CI pipeline. A chain-aware scheduler could have given higher priority to the base (gating seventeen downstream pipelines) and lower priority to any single descendant (gating none until the base merges), turning hours of serialised wait into parallel queueing once the base passes. The principle generalises: OpenStack chain members run 10–23 CI/CD jobs per position against fewer than two for solo changes (Finding #2.3), so wasted resources on superseded patchsets are a recurring property of strict-gating regimes.

Finally, a chain’s dependency graph is not fixed during review: 33.5% of chain members restructure before merging, most often by detaching from an in-review parent. Tools that read the graph once and assume it stable (reviewer recommendation, CI scheduling heuristics, dashboards) produce results based on a graph no longer holding in about one-third of cases. A practical mitigation is to recompute chain-dependent state on every patchset upload, not only at chain creation; the long-tail timing in Table 6 further suggests chain-aware tooling should preserve coordination state across dormant periods rather than discarding context after fixed timeouts.

7 Threats to Validity

Construct validity. We measure review effort (latency, patchset count, comment count) rather than review quality; the foundation effect is therefore framed as propagation of review activity, not of code quality. Chain detection relies on parent-SHA matching, validated against Gerrit’s Relation Chain panel, with the operations/puppet auto-generated deep chain excluded from aggregates and chain evolution distinguishing true structural changes from trivial rebases (same Change-Id). Human-vs-bot distinction in discussion and CI/CD

metrics relies on a per-project allowlist; while this convention is standard in Gerrit-based MCR research [29, 46], residual misclassification at the long tail of commenters cannot be excluded. The Nova patchset-level data was extracted through a replication package rather than our own pipeline; we verified that the same Change-Id-based exclusion was applied, and removing Nova does not qualitatively change the results.

Internal validity. We compare chain members and solo changes within size strata but cannot exclude task complexity as an unmeasured confounder, so the latency overhead is treated as an association rather than a causal effect [5]. We do not apply multiple-comparison corrections to per-project tests because projects are independent observational units, and the consistency of sign and magnitude across the corpus is the primary finding.

External validity. The corpus spans three open-source Gerrit ecosystems, dominated by infrastructure and platform software. Industrial settings, other platforms (GitHub [48], Phabricator, Graphite), and different Gerrit submit types may exhibit different dynamics. Replication on those platforms is a natural next step.

8 Conclusion

We characterized 29,580 relation chains across 401,256 reviewed changes in three open-source Gerrit ecosystems. Chains are prevalent (5–49% of changes, increasing in 14 of 15 projects) and impose measurable coordination costs: members take a median of $2.6\times$ longer to merge than solo changes, exhibit a *CI amplification effect* under strict gating, and place *synchronization bottlenecks* on middle members. The *foundation effect* (Spearman $\rho = 0.43\text{--}0.61$) shows base-change review activity propagates to descendants, and *dynamic dependency restructuring* reshapes one third of chain members mid-review, sometimes across multi-year intermittent timespans.

These regularities suggest a lever for chain-aware tooling: invest review attention in base changes, and the benefit propagates through the chain. Future work should evaluate targeted interventions (base-review checkpoints, chain-aware reviewer assignment [40, 31], dependency-graph-aware CI scheduling) and replicate the analyses on platforms beyond Gerrit (Phabricator, Graphite, GitHub stacked PRs).

Data Availability Statement

To facilitate replication and extension, we share data and scripts in our replication package [1]. The package contains raw Gerrit REST API outputs for the 15 studied projects, processed per-change and per-chain CSVs used as inputs to all analyses, and Jupyter notebooks that reproduce every figure and table reported in this paper, organised by research question. A README documents environment setup and provides a finding-to-cell mapping from each result in Sections 5.1–5.3 to the specific notebook cell that produced it.

References

- 1 Replication package for our study. Figshare, 2026. Available from: <https://figshare.com/s/9ad43e04f03c7b8f7d85>.
- 2 Ali Arabat et al. An ML-based approach to predicting software change dependencies: Insights from an empirical study on OpenStack. *arXiv preprint*, 2025. [arXiv:2508.05034](https://arxiv.org/abs/2508.05034).

- 3 Ali Arabat and Mohammed Sayagh. An empirical study on cross-component dependent changes: A case study on the components of OpenStack. *Empirical Software Engineering*, 29(5):109, 2024. doi:10.1007/s10664-024-10488-y.
- 4 Andrea Arcuri and Lionel Briand. A practical guide for using statistical tests to assess randomized algorithms in software engineering. In *Proceedings of the 33rd International Conference on Software Engineering (ICSE)*, pages 1–10, 2011. doi:10.1145/1985793.1985795.
- 5 Andrea Arcuri and Lionel Briand. A hitchhiker’s guide to statistical tests for assessing randomized algorithms in software engineering. *Software Testing, Verification and Reliability*, 24(3):219–250, 2014. doi:10.1002/STVR.1486.
- 6 Alberto Bacchelli and Christian Bird. Expectations, outcomes, and challenges of modern code review. In *Proc. 35th International Conference on Software Engineering (ICSE)*, pages 712–721, 2013. doi:10.1109/ICSE.2013.6606617.
- 7 Mike Barnett, Christian Bird, João Brunet, and Shuvendu K Lahiri. Helping developers help themselves: Automatic decomposition of code review changesets. In *Proc. 37th International Conference on Software Engineering (ICSE)*, pages 134–144, 2015. doi:10.1109/ICSE.2015.35.
- 8 Olga Baysal, Oleksii Kononenko, Reid Holmes, and Michael W Godfrey. Investigating technical and non-technical factors influencing modern code review. *Empirical Software Engineering*, 21(3):932–959, 2016. doi:10.1007/S10664-015-9366-8.
- 9 Nathan Cassee, Bogdan Vasilescu, and Alexander Serebrenik. The silent helper: The impact of continuous integration on code reviews. In *IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 423–434, 2020. doi:10.1109/SANER48275.2020.9054818.
- 10 Moataz Chouchen and Ali Ouni. A multi-objective effort-aware approach for early code review prediction and prioritization. *Empirical Software Engineering*, 29(1):29, 2024. doi:10.1007/S10664-023-10431-7.
- 11 Moataz Chouchen, Ali Ouni, and Mohamed Wiem Mkaouer. MultiCR: Predicting merged and abandoned code changes in modern code review using multi-objective search. *ACM Transactions on Software Engineering and Methodology*, 33(8):1–44, 2024. doi:10.1145/3680472.
- 12 Moataz Chouchen, Ali Ouni, Jefferson Olongo, and Mohamed Wiem Mkaouer. Learning to predict code review completion time in modern code review. *Empirical Software Engineering*, 28(4):82, 2023. doi:10.1007/S10664-023-10300-3.
- 13 Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993.
- 14 William Jay Conover. *Practical nonparametric statistics*. John Wiley & Sons, 3rd edition, 1999.
- 15 Mattia di Biase, Ayushi Rastogi, and Alberto Bacchelli. The effects of change decomposition on code review—a controlled experiment. *PeerJ Computer Science*, 5:e193, 2019. doi:10.7717/PEERJ-CS.193.
- 16 Harald Gall, Karin Hajek, and Mehdi Jazayeri. Detection of logical coupling based on product release history. In *Proc. International Conference on Software Maintenance (ICSM)*, pages 190–198, 1998. doi:10.1109/ICSM.1998.738508.
- 17 Gerrit Project. Gerrit code review: Changes, 2024. URL: <https://gerrit-review.googlesource.com/Documentation/concept-changes.html>.
- 18 Gerrit Project. Gerrit code review: User review ui – relation chain, 2024. URL: <https://gerrit-review.googlesource.com/Documentation/user-review-ui.html#related-changes>.
- 19 GitHub. Github stacked PRs, 2026. Native stacked pull request support, currently in private preview. URL: <https://github.blog/changelog/>.
- 20 Graphite. Why stacked diffs lead to better code, 2024. URL: <https://graphite.dev/guides/stacked-diffs>.

- 21 Qi Guo, Junming Cao, Xiaofei Xie, Shangqing Liu, Xiaohong Li, Bihuan Chen, and Xin Peng. Exploring the potential of ChatGPT in automated code refinement: An empirical study. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering (ICSE)*, 2024.
- 22 Ahmed E Hassan. Predicting faults using the complexity of code changes. In *Proceedings of the 31st International Conference on Software Engineering (ICSE)*, pages 78–88, 2009. doi:10.1109/ICSE.2009.5070510.
- 23 Toshiki Hirao, Shane McIntosh, Akinori Ihara, and Kenichi Matsumoto. The review linkage graph for code review analytics: a recovery approach and empirical study. In *Proc. 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pages 578–589, 2019. doi:10.1145/3338906.3338949.
- 24 Oleksii Kononenko, Olga Baysal, and Michael W Godfrey. Code review quality: how developers see it. In *Proc. 38th International Conference on Software Engineering (ICSE)*, pages 1028–1038, 2016. doi:10.1145/2884781.2884840.
- 25 Zhiyu Li, Shuai Lu, Daya Guo, Nan Duan, Shailesh Jannu, Grant Jenks, Deep Majumder, Jared Green, Alexey Svyatkovskiy, Shengyu Fu, et al. Automating code review activities by large-scale pre-training. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pages 1035–1047, 2022.
- 26 Laura MacLeod, Michaela Greiler, Margaret-Anne Storey, Christian Bird, and Jacek Czerwinka. Code reviewing in the trenches: Challenges and best practices. *IEEE Software*, 35(4):34–42, 2017. doi:10.1109/MS.2017.265100500.
- 27 Rungroj Maipradit, Dong Wang, Patanamon Thongtanunam, Raula Gaikovina Kula, Yasutaka Kamei, and Shane McIntosh. Repeated builds during code review: An empirical study of the OpenStack community. In *Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 153–165, 2023. doi:10.1109/ASE56229.2023.00030.
- 28 Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.
- 29 Shane McIntosh, Yasutaka Kamei, Bram Adams, and Ahmed E Hassan. An empirical study of the impact of modern code review practices on software quality. *Empirical Software Engineering*, 21(5):2146–2189, 2016. doi:10.1007/S10664-015-9381-9.
- 30 Meta. ghstack: Submit stacked diffs to GitHub. URL: <https://github.com/ezyang/ghstack>.
- 31 Ehsan Mirsaeedi and Peter C Rigby. Mitigating turnover with code review recommendation: Balancing expertise, workload, and knowledge distribution. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE)*, pages 1183–1195, 2020. doi:10.1145/3377811.3380335.
- 32 Phacility. Phabricator documentation: Differential, 2024. URL: <https://secure.phabricator.com/book/phabricator/article/differential/>.
- 33 Anand Ram, Anand Ashok Sawant, Marco Castelluccio, and Alberto Bacchelli. What makes a code change easier to review: an empirical investigation on code change reviewability. In *Proc. 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pages 201–212, 2018.
- 34 Peter C Rigby and Christian Bird. Convergent contemporary software peer review practices. In *Proc. 9th Joint Meeting on Foundations of Software Engineering (ESEC/FSE)*, pages 202–212, 2013. doi:10.1145/2491411.2491444.
- 35 Peter C Rigby, Daniel M German, and Margaret-Anne Storey. Open source software peer review practices: a case study of the Apache server. In *Proc. 30th International Conference on Software Engineering (ICSE)*, pages 541–550, 2008.
- 36 Jeanine Romano, Jeffrey D Kromrey, Jesse Coraggio, and Jeanine Skowronek. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen’s d for evaluating group differences on the NSSE and other surveys? In *Annual Meeting of the Florida Association of Institutional Research*, pages 1–33, 2006.

- 37 Pranab Kumar Sen. Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association*, 63(324):1379–1389, 1968.
- 38 Patanamon Thongtanunam, Shane McIntosh, Ahmed E Hassan, and Hajimu Iida. Revisiting code ownership and its relationship with software quality in the scope of modern code review. In *Proceedings of the 38th International Conference on Software Engineering (ICSE)*, pages 1039–1050, 2016. doi:10.1145/2884781.2884852.
- 39 Patanamon Thongtanunam, Shane McIntosh, Ahmed E Hassan, and Hajimu Iida. Review participation in modern code review: An empirical study of the Android, Qt, and OpenStack projects. *Empirical Software Engineering*, 22(2):768–817, 2017. doi:10.1007/S10664-016-9452-6.
- 40 Patanamon Thongtanunam, Chakkrit Tantithamthavorn, Raula Gaikovina Kula, Norihiro Yoshida, Hajimu Iida, and Kenichi Matsumoto. Who should review my code? a file location-based code-reviewer recommendation approach for modern code review. In *22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, pages 141–150, 2015.
- 41 Rosalia Tufano, Simone Masiero, Antonio Mastropaolo, Luca Pascarella, Denys Poshyvanyk, and Gabriele Bavota. Using pre-trained models to boost code review automation. In *Proceedings of the 44th International Conference on Software Engineering (ICSE)*, pages 2291–2302, 2022. doi:10.1145/3510003.3510621.
- 42 Bogdan Vasilescu, Yue Yu, Huaimin Wang, Premkumar Devanbu, and Vladimir Filkov. Quality and productivity outcomes relating to continuous integration in GitHub. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering (ESEC/FSE)*, pages 805–816, 2015. doi:10.1145/2786805.2786850.
- 43 Dong Wang et al. An exploration of cross-patch collaborations via patch linkage in OpenStack. In *arXiv preprint arXiv:2211.15007*, 2022.
- 44 Dong Wang, Raula Gaikovina Kula, Takashi Ishio, and Kenichi Matsumoto. Automatic patch linkage detection in code review using textual content and file location features. *Information and Software Technology*, 139:106637, 2021. doi:10.1016/J.INFSOF.2021.106637.
- 45 Peter Weißgerber, Daniel Neu, and Stephan Diehl. Small patches get in! In *Proc. 2008 International Working Conference on Mining Software Repositories (MSR)*, pages 67–76, 2008. doi:10.1145/1370750.1370767.
- 46 Mairieli Wessel, Alexander Serebrenik, Igor Wiese, Igor Steinmacher, and Marco A Gerosa. Quality gatekeepers: Investigating the effects of code review bots on pull request activities. *Empirical Software Engineering*, 27(5):108, 2022. doi:10.1007/S10664-022-10130-9.
- 47 Yue Yu, Huaimin Wang, Vladimir Filkov, Premkumar Devanbu, and Bogdan Vasilescu. Wait for it: Determinants of pull request evaluation latency on GitHub. In *Proceedings of the 12th Working Conference on Mining Software Repositories (MSR)*, pages 367–371, 2015. doi:10.1109/MSR.2015.42.
- 48 Xunhui Zhang, Yue Yu, Georgios Gousios, and Ayushi Rastogi. Pull request decision explained: An empirical overview. *IEEE Transactions on Software Engineering*, 49(2):849–868, 2022.
- 49 Thomas Zimmermann, Andreas Zeller, Peter Weissgerber, and Stephan Diehl. Mining version histories to guide software changes. *IEEE Transactions on Software Engineering*, 31(6):429–445, 2005. doi:10.1109/TSE.2005.72.