When Do You Repeat Yourself? Voices from the Trenches of Linux Kernel Maintainers on Code Duplication

Luan Arcanjo
David Tadokoro
luanicaro@usp.br
davidbtadokoro@ime.usp.br
University of São Paulo
São Paulo, Brazil

Marcelo Spessoto
Rafael Passos
marcelomspessoto@usp.br
rcpassos@ime.usp.br
University of São Paulo
São Paulo, Brazil

Paulo Meirelles
paulormm@ime.usp.br
University of São Paulo
São Paulo, Brazil

Abstract

The Don't Repeat Yourself (DRY) principle is central to software maintainability, but empirical studies challenge its rigid use, describing beneficial cases of duplication. However, these rely on retrospective analyses, leaving a gap in understanding real-time decision-making and socio-technical dynamics. This paper presents an ethnographic study on how the Linux kernel community manages duplication debt via deduplication contributions. Using a clone detection tool called ArKanjo, we conducted a multimethod ethnographic study: first as a complete participant submitting patches to AMDGPU, then as a participant-as-observer mentoring 23 newcomers contributing to AMDGPU and IIO. Analysis of patch reviews suggests maintainers can tolerate duplication to accommodate driver-forking (T1), prioritize readability (T2), reduce integration overhead (T3), and preserve performance (T4). Our findings demonstrate that managing duplication in Linux is a nuanced process in which trade-offs among maintainability, clarity, and practicality outweigh dogmatic adherence to the DRY principle.

CCS Concepts

• **Software and its engineering → Open source model**; • **Human-centered computing → Open source software**.

Keywords

Linux kernel, Code Quality, Deduplication, Maintainers

ACM Reference Format:
Luan Arcanjo, David Tadokoro, Marcelo Spessoto, Rafael Passos, and Paulo Meirelles. 2026. When Do You Repeat Yourself? Voices from the Trenches of Linux Kernel Maintainers on Code Duplication. In *International Conference on Technical Debt (TechDebt '26), April 12–13, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3794915.3795783

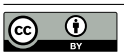
1 Introduction

Having tens or hundreds of developers working together on a software product is a complex and demanding task. Similar or even exact copies are a frequent and often unavoidable outcome. Regarded as a bad smell [5] and a major source of technical debt, *code duplication* (or *code cloning*) can negatively impact a project's maintainability, as any change to a repeated code segment (to fix a bug, for instance) requires replication across all other copies. Developers must therefore remain constantly vigilant to synchronize different parts of the codebase, a seemingly simple yet highly error-prone responsibility [9]. Removing duplicated code (*deduplication*) demands great care to preserve the original behavior, with minor mistakes possibly leading to serious problems in the future.

A widespread dogma among developers is the *Don't Repeat Yourself* (DRY) principle [8], which asserts that to develop reliable and maintainable software, "*every piece of knowledge must have a single, unambiguous, authoritative representation within a system.*" Through static program analysis and mining software repository techniques, some empirical studies conducted on living real-world projects have challenged the DRY principle, showing that duplication can be advantageous depending on the context and evolution of the clones [7, 11, 15, 21]. Nevertheless, these types of analysis primarily operate in hindsight; identifying, cataloguing, and drawing conclusions from the already made and final decisions regarding technical debt management, which, in turn, can obscure **the reasons and the whole story behind accepting certain clone debts or actively blocking repayments in complex projects**. This work aims to capture that process as it unfolds.

To investigate these decision-making processes in a realistic setting, we turn to one of the largest and most influential software systems ever developed: the Linux kernel, a foundational *Free/Libre and Open Source Software* (FLOSS) project with more than 27 million lines of code, excluding non-programming code (verifiable through our replication package; see Section 7), and over 2,000 developers involved in version 6.17 [3]. Although it is known that some Linux device drivers deliberately duplicate code [11], **it can be hard to determine why a specific clone exists**, even when analyzing upstream commits in the mainline and subsystems. One could argue that the debt interest is too low and the debt principal is too high, or that there is a motive for accepting the clone unrelated to technical debt, or that simply no reasonable deduplication proposal has ever been submitted. We are left to speculate (in a well-informed manner) about developers' debt management practices and the reasoning behind each commit we analyze.

Leveraging a function-level clone detection tool we developed specifically for the context of Linux, called *ArKanjo* [1], we conducted a multimethod ethnographic study regarding code duplication. Ethnography is a well-established qualitative method for



TechDebt '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2485-5/2026/04
https://doi.org/10.1145/3794915.3795783

understanding people, their cultures, and work practices [4]. It provides insights into community members' values, beliefs, and practices [18]. In this study, **we detected duplications and submitted deduplications to two Linux subsystems.** Through interactions with maintainers, we gained insights into how the Linux community perceives code quality related to code clones in the project's "development trenches". Contrary to the DRY principle, our findings confirm that, in some contexts, maintainers can tolerate duplication to improve readability. **We also identified reasons coupled to enhancing performance and avoiding integration overhead**. This empirical work also offers a distinct approach to researching technical debt management practices related to code duplication in FLOSS projects, beyond indicating that deduplication contributions are a viable entryway for newcomers in the Linux kernel.

2 Related Work

The works by Kamiya et al. [10] and Li et al. [13] proposed techniques to identify duplications, using the Linux codebase as a case study. Kapser and Godfrey [11] discussed patterns of cloning used in software, including situations where this could be positive, and Rahman et al. [15] contested the notion of clones as a bad smell. Tornhill [21] claims that clones evolving independently might not be problematic, whilst Juergens et al. [9] argue that inconsistent changes to clones that should evolve together are what primarily lead to faults. Wang and Godfrey [22] explore the use of cloning as a development technique for the SCSI drivers in the Linux Kernel, where similar devices had drivers born from code cloning. Harder [7] acknowledges the harmful case of cloned bugs but provides empirical evidence contradicting popular beliefs regarding the adverse effects of clones. It did not provide clear guidance for practitioners, leaving the way open for works such as this one.

Our study provides additional empirical evidence and a distinct methodological perspective for researchers and practitioners interested in technical debt, code quality, and socio-technical practices within the Linux kernel project. **We obtained key insights by analyzing maintainers' reasoning when they accepted or rejected submitted changes (*patches*)**. In this respect, we make a new contribution to a long-standing discussion among software engineering researchers about when and how duplicate code matters.

3 Research Strategy

First, we needed a tool for detecting code duplication suitable for the Linux context. After reviewing existing static analysis tools such as *PMD*, *Duplo*, *Simian*, and *CodeClimate* (the analysis of the candidate tools is available in our replication package; see Section 7), we identified practical limitations when applying them to specific Linux kernel subsystems. Many were commercially licensed, used only token-based matching, or, like some FLOSS alternatives, were limited to file-scope comparisons, missing intra-file duplications. To address these constraints, we developed ArKanjo, a custom command-line tool for detecting duplicated functions in large C codebases. Released under LGPLv3, it uses *Term Frequency-Inverse Document Frequency* (TF-IDF) vector embeddings and cosine similarity [16, 17] that identify the four clone types described in the literature (Type-1 to Type-4) [2]; at this moment, it was validated and works more accurately with Type-1 and Type-2 clones [1]. This lightweight Command Line Interface (CLI) prioritizes maintainability, keeping our focus on code analysis and community interactions rather than tool development.

The Linux kernel is organized into subsystems like the process scheduler, memory management, and device drivers, each typically overseen by a dedicated maintainer or team. Our study focuses on two subsystems representing different contribution contexts. The first is the *AMDGPU DRM drivers* subsystem, which natively enables AMD GPU functionality in Linux. This substantial subsystem contains over 391,000 lines of code across more than 1,089 files. We chose it because, similar to the SCSI subsystem studied by Wang and Godfrey [22], AMDGPU has a multi-layered architecture with generic higher layers and lower layers comprising hardware-specific drivers implementing the same basic functionality. The second subsystem is the *Industrial Input/Output* (IIO) subsystem, chosen because it is considered an accessible entry point for new contributors. IIO provides support for analog-to-digital or digital-to-analog converters with over 281,000 lines of code in more than 755 files.

Using ArKanjo, we assessed the viability of reducing the identified duplications through a two-phase ethnographic study employing participant observation methodology [6]. In the first phase, adopting the role of *complete participant* [6], we engaged directly with the Linux kernel community by submitting deduplication patches to the AMDGPU drivers. In the second phase, shifting to the role of *participant-as-observer* [6], we introduced a similar activity in a university course, in which students were tasked with proposing patches for the AMDGPU or IIO subsystems. Throughout this process, we provided guidance and mentorship to the students as they interacted with the community.

3.1 Complete Participant Observation Study

We first conducted a complete participant observation experiment, in which the main author (a graduate student) served as a first-time contributor to the AMDGPU drivers, collecting artifacts of his experience along the way. We ran the ArKanjo tool on the AMDGPU codebase and manually analyzed the largest duplications by line count, which were Type-1 or Type-2, and identified one pair we judged promising to try to deduplicate. In the files of the duplicated function pair, we observed multiple other Type-1 and Type-2 clones. Thus, we proposed a simple systematic approach to mitigate all the duplicate functions in this context, not just the initial function pairs. After approaching this refactoring with a systematic strategy, we sent a patch to the AMDGPU drivers mailing list for the maintainers' feedback, while documenting the process, interactions, and impressions (also available in our replication package; see Section 7).

In C programming, communication between source files is achieved by creating header files that specify libraries [12]. Since the Linux kernel is primarily written in C, our systematic strategy consisted of eliminating code clones by consolidating them into a single library, thereby replacing duplicate instances across the codebase. For each flagged function, we used the tool to locate all duplications of that function in the AMDGPU codebase, since it may have been replicated beyond the two occurrences initially

detected. This strategy resulted in a collection of functions and corresponding duplication occurrences in code files.

To identify shared code more effectively, we extended this approach to search for other common functions across all collected files. We then applied specific refactoring methods to each shared function. If the functions were identical across files, we removed the duplicates and created a single function in the library. If modifications existed, we applied case-specific refactoring.

3.2 Participant-as-Observer Observation Study

To broaden our study and observations, we adopted a participant-as-observer approach, involving students as FLOSS project newcomers to make practical contributions to the Linux kernel. This approach formed the core activity of a university course [20], in which we guided students to remove code duplications in the Linux kernel as one possible pathway for contributing to the project in the first phase of the course.

The 2025 course offering had 37 students (25 undergraduate and 12 graduate), who were asked to form groups of two or three members, but graduate students could work individually [20]. To assist students in this task, we (a professor and three more experienced graduate students) prepared several alternatives for contributing to the Linux kernel, including simpler contributions such as cleaning coding style issues. Regarding deduplication, we specified two options in the Linux kernel.

Deduplication Option 1. We ran ArKanjo on the IIO subsystem to generate a list of duplicated function pairs, then curated it to highlight actionable cases for students. Specifically, we (i) kept only pairs within the same file, since cross-driver duplications could involve multiple maintainers or distinct interfaces; (ii) ignored very short duplications; (iii) considered only Type-1 and Type-2 clones, since semantical Type-3 and Type-4 present higher complexity in refactoring them; and (iv) manually ranked the remaining pairs, adding annotations to guide students toward entries likely to be accepted. In this option, students were tasked with selecting a recommended entry, devising a way to remove the duplication, and submitting their patches to the IIO mailing list for review.

Deduplication Option 2. We offered an experience similar to our initial complete participant observation, where students managed the entire workflow: running the tool, analyzing results to identify potential deduplications, creating a patch, and submitting it to the driver's maintainers. This option provided more freedom of choice but made the students' task more complex.

Of the students who chose to work on the duplication issues, 23 (16 undergraduate and 7 graduate) formed 11 groups to pursue the first alternative. One graduate student opted for the second alternative. None of these students had prior experience contributing to the Linux kernel, making this their first attempt to submit a patch as newcomers to the project.

It is important to stress that, **no matter the chosen approach, we supported all groups by providing review cycles on their contributions before submission**, to prepare students for the interactions with the maintainers and alleviate rough, but natural, newcomers' mistakes [20]. After the submission, we helped students understand the feedback from maintainers and gave directions on how to move forward. This close, yet indirect, participation through intimate mentoring characterizes the second phase as our participant-as-observer observation.

For this second phase, our primary data source was the mailing list threads, where all development occurred (also available in our replication package; Section 7). Using blog posts, the groups documented their experiences submitting a patch to the Linux kernel, and we also leveraged those to complement our observations.

4 Results and Discussion

In the **complete participant observation** study (first phase), we sent the first revision of the patch on August 9, 2024, and a resend followed on October 9, 2024, due to a lack of initial feedback. Nevertheless, for the first two versions, the maintainers' response requested minor changes for coding style, license use, and alignment with existing conventions. In the third version, the feedback was to move the new generic library functions into an existing file rather than creating a new one. Then, the fourth version was accepted and integrated into the kernel on February 25, 2025, removing 406 lines in total and impacting three generations of drivers.

Since we experienced a considerable delay in the review process, we directly contacted a maintainer to understand the cause. From this conversation, we understood that it is commonplace for drivers to be cloned (sometimes entirely) to provide a solid base and enhance independence between them, allowing developers to make changes to a specific one without needing to check for regressions on others. This practice is very similar to that detected in the SCSI subsystem by Wang and Godfrey [22], agrees with Tornhill [21] regarding clones evolving independently, and is also the hardware variation of the forking pattern defined by Kapser and Godfrey [11]. Even so, this driver duplication is a *Self-Admitted Technical Debt* (SATD) [14] listed as a welcomed contribution in the official documentation of the Display Core component[1]. In this framing, we can attribute the delay to the change impacting multiple drivers, which resulted in greater principal on the maintainer's side to ensure the repayment was stable. Prioritization of other repayments may also have contributed to the delay. With that in mind, this complete participant observation produced our initial takeaway:

> **Takeaway T1 (Driver Forking):** *Driver developers do not necessarily view duplicated code negatively, often consciously cloning entire drivers for many reasons; however, sometimes such duplications are explicitly acknowledged as technical debt, and their repayments are accepted when perceived as safe and stable.*

Our initial observations indicate that, although driver forking has been previously reported in large-scale systems [11, 22], our experience in the Linux kernel shows that removing such duplication can be accepted in practice, pointing to opportunities for further collaboration and empirical studies on technical debt management in this context.

To further expand T1 and possibly derive other takeaways, we examined the contributions made through our **participant-as-observer observation**, substantiating them with quotes from the maintainers that emerged during the review process[2]. This broader

[1] https://www.kernel.org/doc/html/v6.17/gpu/amdgpu/display/display-contributing.html#reduce-code-duplication.

[2] We **do not claim** that these quotes fully represent the maintainers' stance on code duplication or any technical debt management matter.

analysis aimed to capture diverse maintainer perspectives and contextual factors influencing their decisions.

Table 1: Summary of newcomer deduplication contributions.

GID	SS	SML	Patch Status	Takeaway	Diff
0	AMD	100%	Accepted (v4)	T1	+114/-520
1	AMD	90%	Accepted (v2)	T1	+90/-489
	IIO	100%	Dropped (v1)	T1	+2/-12
2	IIO	100%	Dropped (v2)	T2, T3	+7/-14
3	IIO	100%	Dropped (v1)	T2	+23/-27
4	IIO	100%	Accepted (v1)	T2	+12/-30
5	IIO	100%	Accepted (v3)	T2	+2/-7
6	IIO	90%	Accepted (v1)	T2	+17/-70
7	IIO	90%	Dropped (v1)	T2	+95/-92
8	IIO	90%	Accepted (v3)	-	+20/-36
9	IIO	90%	Accepted (v4)	T4	+20/-16
10	IIO	90%	Accepted (v6)	T4	+149/-243
11	IIO	90%	Dropped (v1)	T2, T4	+32/-58

Table 1 summarizes all the contributions sent by the main author and student groups (GID)[3]. It contains the information about the chosen subsystem (SS) and the similarity threshold (SML). Each entry also has the patch status, takeaways that may have surfaced from the interaction, and a code differential (Diff) corresponding to the final patch versions before it was dropped or accepted into the code base. The code differential for a set of patches corresponds to the sum of added and removed lines of each patch.

Groups 1 to 11 all sent one patch to IIO from the deduplication option 1 described in Section 3, yet Group 1 also submitted one to AMDGPU using the deduplication option 2.

The AMDGPU patch of Group 1 was similar to the earlier complete participant patch, but was merged faster (within 10 days). This expands T1, highlighting how developers deliberately fork drivers as standard practice yet recognize the resulting duplication as debt and accept repayments. The IIO patch of Group 1 was rejected and targeted a driver using the *Register Map* API, which requires a `regmap_config` struct. It incorrectly inlined two callback functions returning false in the struct (which require function pointers, not booleans), as ambiguously noted by the maintainer: *"Take another look at what you are doing here."* This reflects the API/library protocol templating pattern described by Kapser and Godfrey [11], which our analysis also traces to driver forking, further reaffirming T1.

Groups 2 and 5 both refactored similar predicate functions that were logical negations of each other. Although initially correct, both solutions underwent significant changes driven by maintainer concerns about readability. For Group 2, a maintainer said that *"[...] the naming as* `_reg_check()` *is not helpful as it doesn't indicate anything specific is being checked."*, while for Group 5, the feedback explicitly highlighted the precedence of readability over deduplications: *"I think the old code is more readable than hiding the values in a macro even if it is duplicating a few lines of code."* Nonetheless, after both groups incorporated their respective feedback from the maintainers, only Group 5 had its patch accepted. Even though the deduplication of Group 3 (dropped) was not similar to that of Group 2 (dropped) and Group 5 (accepted) in terms of implementation (it used the parameterize method), it also brought forth the readability factor: *"In my view this isn't a significant enough reduction to justify the more complex code."*

Groups 4, 6, 7, and 11 used a parameterized approach with a new generic function parameter. The same maintainer reviewed all four, but only Groups 4 and 6 had their patches accepted. For Group 4, we think this was because it applied to a straightforward iterative algorithm and also fixed a bug, helping its approval, whereas we were not able to identify a solid, explicit reason for accepting the patch for Group 6. Although deduplication does not always reduce lines, the patch from Group 7 added lines, which the maintainer cited as a drawback. Still, the real reason for the rejection was that the deduplication scattered parts of a complex code that resided close together before: *"Wrapping this up doesn't provide any real advantage, requiring as it does the reviewer to look at this function AND where the value is set rather than seeing them in one place."* Needing to jump to a function definition (to see its behavior) or a function call (to see its concrete parameters) to understand code is a cognitively demanding task for developers [19]. The reason for rejecting the patch from Group 11 was that the deduplication lowered readability, as the maintainer was *"[...] not sure the code reduction is sufficiently to cover the resulting loss of readability."*.

Understandability can motivate cloning, to the extent that refactoring may harm conceptual cohesiveness [11]. Empirical observations support this, as it led to dropping patches from Groups 2, 3, 7, and 11. Meanwhile, accepted patches from Groups 4, 5, and 6 may reduce readability based on similar maintainer arguments from the dropped patches. As with T1, these contrasting views on duplication and readability reveal nuance and highlight a need for further research. Thus, we propose our second takeaway:

> **Takeaway T2 (Readability):** *Maintainers prioritized readability over removing duplication across many contexts. Accepting debt from duplication may be worth it if it results in easier-to-read, more understandable code. Nevertheless, the criteria used to accept or reject similar repayments that affect readability require deeper investigation.*

Going back to Group 2, maintainers questioned the merit of the proposed patch. While the previous interactions highlighted how maintainers often favor readability over strict code deduplication, this case added a new dimension to our understanding: the *cost-benefit* evaluation of integrating proposed changes. A maintainer pointed out that the effort required to propagate such modifications after acceptance (i.e., during the upstreaming process) could outweigh the repayment benefits: "*[...] such patches might not worth it since the proposed improvement is very small (and questionable) while the upstreaming process still requires some effort.*" From this feedback, we crafted our third takeaway:

> **Takeaway T3 (Integration Overhead):** *Maintainers may decline to accept a duplication repayment if it is not considered impactful enough, because simply integrating it into the Linux upstream incurs considerable principal.*

In the interaction of Group 11, there was also a hint that performance could be considered when evaluating the worth of a

[3]Since the first author was also a newcomer, the first phase experience was included as Group 0.

deduplication, whilst reaffirming that readability was a priority in that context: *"I'd be slightly interested to see the optimized output of the two approaches, but this is far from a high performance path so we care a lot more about readability here."* Nonetheless, this notion surfaced aggressively with Group 10, where the contribution could completely remove both duplicated functions using the inline method by leveraging helpers; however, one of the functions had a clone instance that, if eliminated, would result in a significant performance impact: *"Note that is not an appropriate change for the large reads though [...] This is the one case were* `spi_write_then_read()` *[the helper] is probably not appropriate due to the large buffers that are potentially involved."* Group 10 complied to the suggestion, keeping one of the functions for this single occurence, and had its patch accepted.

Groups 8 and 9 applied the extract method with helpers for deduplication. Group 8 refined its patch using a library helper function suggested by the maintainer, resulting in smooth acceptance. While not directly linked to our takeaways, this case reveals latent duplication repayments: instances of technical debt that remain unaddressed until external contributions prompt action. Group 9 followed a similar path but performance concerns were raised: one maintainer claimed that *"Even though there is less code repetition, we now have an extra comparison [...]"*, and another endorsed this in version three *'There is always a balance/trade-off between modularity and execution speed. I agree with* `anonymous-maintainer`*'s reply in the first patch [...]"*. Notwithstanding, after doing the necessary corrections, the patch was accepted in version four.

The interactions from Groups 9 (accepted), 10 (accepted), and 11 (dropped), coupled with the low-level development context of Linux, produced our fourth takeaway:

> **Takeaway T4 (Performance):** *Interactions with maintainers hinted that duplications can be forgiven if they improve performance in contexts where resources are limited. In these scenarios, avoiding performance debt is prioritized.*

In summary, as previously mentioned, T1 and T2 have been reported in the literature, but **our close ethnographic observation also yields differing, sometimes contradictory, perspectives that warrant further investigation**. Additionally, takeaways T3 and T4 go beyond and **offer nuanced insights that seem specific to debt management decisions that Linux maintainers practice**.

Moreover, another engaging aspect stemming from this work: deduplication offers interesting opportunities in the Linux kernel and can be an excellent approach for newcomers.

Finally, in terms of numbers, the 13 contributions included: (i) a total of 585 lines added and 1,626 lines removed; (ii) 8 contributions were accepted, yielding a success rate of 62%; (iii) for accepted contributions, the total code diff was +255/-1,152, with an average of +31.9/-144.0 lines added/removed; and (iv) for rejected contributions, the total code diff was +181/-231, averaging +36.2/-46.2 lines added/removed.

5 Threats to Validity

We outline limitations using standard validity categories. First (***Internal Validity***), selection bias may affect findings. Targeting large duplications in AMDGPU likely captured complex, conservative cases, while curating simpler duplications for participants may have oversimplified challenges and favored easier patches. Results may not generalize to medium-complexity or inter-file duplications. Second (***External Validity***), findings are based on two Linux kernel subsystems (AMDGPU and IIO) and 23 student contributors. The kernel is a unique FLOSS project with specific norms so that maintainer perspectives may vary. The lack of previous kernel experience from participants also limits generalization to experienced contributors. Third (***Construct Validity***), we did not consider Type-3 and Type-4 duplications, which may limit the analyses regarding more semantically similar duplications. Fourth (***Conclusion Validity***), the ethnographic approach provides qualitative insight but not statistical generalization. Conclusions are drawn from 13 patch interactions. In this sense, a larger sample would strengthen robustness.

6 Concluding Remarks

This study, combining complete participant and participant-as-observer observations, provided a realistic view of the opportunities and challenges in repaying duplication debts in the Linux kernel. We employed an ethnographic approach to engage with technical debt literature that challenges the DRY principle, while refining the discussion within the Linux project.

Our first two empirical findings have been discussed in the literature on the benefits of duplication: detecting the driver-forking pattern (T1) and duplication motivated by readability (T2). However, contrasting observations across these contexts highlight significant nuances that warrant more in-depth investigation. The latter two findings highlight project-specific debt management in the Linux kernel: change propagation considerably increases principal (T3), and performance preservation in resource-constrained settings (T4) further justifies tolerating clones.

Without disregarding its limitations, this work shows that **the Linux community has particular technical debt management practices**, carefully weighing the pros and cons of accepting versus repaying duplication debts. Maintainability, clarity, and practical trade-offs shape maintainers' judgments, suggesting that consciously asking *"**When** do you repeat yourself?"* can be better than mindlessly following the *"Don't Repeat Yourself"* principle.

7 Replication Package

All anonymized review threads are available in this Zenodo repository: doi.org/10.5281/zenodo.17503684. Replication resources include the ArKanjo repository (github.com/arkanjo-tool/arkanjo), its documentation (arkanjo-tool.github.io/), and an analysis of candidate tools (github.com/arkanjo-tool/arkanjo/wiki/Evaluation-of-other-duplication-tools). The container for the Linux v6.17 LoC statistic from Section 1 is also provided (github.com/linux-duks/linux-programming-loc).

Acknowledgments

This study is funded by the São Paulo Research Foundation (FAPESP) and the São Paulo State Data Analysis System Foundation (SEADE), under grants 2023/18026-8 and 2025/05395-0.

References

[1] Luan Arcanjo, David Tadokoro, and Paulo Meirelles. 2025. ArKanjo: a tool for detecting function-level Code Duplication in the Linux Kernel. In *DebConf25*, Olivier Barais and Bastien Roucariès (Eds.). IRISA, Brest, France, 3. https://hal.science/hal-05335545

[2] Chang-Feng Chen, Azlan Zain, and Kai-Qing Zhou. 2022. Definition, approaches, and analysis of code duplication detection (2006–2020): a critical review. *Neural Computing and Applications* 34 (08 2022), 1–31. doi:10.1007/s00521-022-07707-2

[3] Jonathan Corbet. 2025. *Development statistics for 6.17.* Linux Weekly News. https://lwn.net/Articles/1038358/ [Online; Last accessed on Jan. 25th, 2026].

[4] W Alex Edmonds and Thomas D Kennedy. 2016. *An applied guide to research designs: Quantitative, qualitative, and mixed methods.* Sage Publications.

[5] Martin Fowler. 2018. *Refactoring: improving the design of existing code* (2 ed.). Addison-Wesley Professional.

[6] Raymond L. Gold. 1958. Roles in Sociological Field Observation. *Social Forces* 36 (1958), 217–223. doi:10.2307/2573808

[7] Jan Harder. 2017. *Software Clones-Guilty Until Proven Innocent?* Logos Verlag Berlin GmbH.

[8] Andrew Hunt and David Thomas. 2000. *The pragmatic programmer: from journeyman to master.* Addison-Wesley Longman Publishing Co., Inc., USA.

[9] Elmar Juergens, Florian Deissenboeck, Benjamin Hummel, and Stefan Wagner. 2009. Do code clones matter?. In *2009 IEEE 31st International Conference on Software Engineering*. 485–495. doi:10.1109/ICSE.2009.5070547

[10] T. Kamiya, S. Kusumoto, and K. Inoue. 2002. CCFinder: a multilinguistic token-based code clone detection system for large scale source code. *IEEE Transactions on Software Engineering* 28, 7 (July 2002), 654–670. doi:10.1109/TSE.2002.1019480

[11] Cory J. Kapser and Michael W. Godfrey. 2008. "Cloning considered harmful" considered harmful: patterns of cloning in software. *Empirical Software Engineering* 13, 6 (Dec. 2008), 645–692. doi:10.1007/s10664-008-9076-6

[12] Brian W Kernighan and Dennis M Ritchie. 1988. *The C programming language.* prentice-Hall.

[13] Z. Li, S. Lu, S. Myagmar, and Y. Zhou. 2006. CP-Miner: finding copy-paste and related bugs in large-scale software code. *IEEE Transactions on Software Engineering* 32, 3 (March 2006), 176–192. doi:10.1109/TSE.2006.28

[14] Aniket Potdar and Emad Shihab. 2014. An Exploratory Study on Self-Admitted Technical Debt. In *2014 IEEE International Conference on Software Maintenance and Evolution*. 91–100. doi:10.1109/ICSME.2014.31

[15] Foyzur Rahman, Christian Bird, and Premkumar Devanbu. 2012. Clones: what is that smell? *Empirical Software Engineering* 17, 4 (01 Aug 2012), 503–530. doi:10.1007/s10664-011-9195-3

[16] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24, 5 (Aug. 1988), 513–523. doi:10.1016/0306-4573(88)90021-0

[17] G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 11 (Nov. 1975), 613–620. doi:10.1145/361219.361220

[18] Helen Sharp, Yvonne Dittrich, and Cleidson R. B. de Souza. 2016. The Role of Ethnographic Studies in Empirical Software Engineering. *IEEE Transactions on Software Engineering* 42, 8 (2016), 786–804. doi:10.1109/TSE.2016.2519887

[19] Jonathan Sillito, Gail C. Murphy, and Kris De Volder. 2006. Questions programmers ask during software evolution tasks. In *Proceedings of the 14th ACM SIGSOFT International Symposium on Foundations of Software Engineering* (Portland, Oregon, USA) *(SIGSOFT '06/FSE-14)*. Association for Computing Machinery, New York, NY, USA, 23–34. doi:10.1145/1181775.1181779

[20] David Tadokoro, Rafael Passos, and Paulo Meirelles. 2025. Guidelines for Boosting Long-Lasting FLOSS Contributors. In *DebConf25*, Olivier Barais and Bastien Roucariès (Eds.). IRISA, Brest, France, 6. https://hal.science/hal-05334509

[21] Adam Tornhill. 2018. *Software Design X-Rays: Fix Technical Debt with Behavioral Code Analysis* (1st ed ed.). The Pragmatic Programmers, LLC.

[22] Wei Wang and Michael W. Godfrey. 2011. A Study of Cloning in the Linux SCSI Drivers. In *2011 IEEE 11th International Working Conference on Source Code Analysis and Manipulation*. 95–104. doi:10.1109/SCAM.2011.17