The Strength of Weak Ties among Open-Source Developers

In the Python Ecosystem, Diversity via GitHub Stars Predicts Innovation Better Than Diversity via Commits

HONGBO FANG, Carnegie Mellon University, University of Chicago, USA
PATRICK PARK, Carnegie Mellon University, USA
JAMES EVANS, The University of Chicago, Santa Fe Institute, USA
JAMES HERBSLEB, Carnegie Mellon University, USA
BOGDAN VASILESCU, Carnegie Mellon University, USA

In a real-world social network, weak ties (reflecting low-intensity, infrequent interactions) act as bridges and connect people to different social circles, giving them access to diverse information and opportunities that are not available within one's immediate, close-knit vicinity. Weak ties can be crucial for creativity and innovation, as they introduce ideas and approaches that people can then combine in novel ways, leading to innovative solutions. Do weak ties facilitate creativity in software in similar ways? This paper suggests that the answer is "yes." Concretely, we study the correlation between developers' knowledge acquisition through three distinct interaction networks on GitHub and the innovativeness of the projects they develop, across over 37,000 Python projects hosted on GitHub. Our findings suggest that the topical diversity of projects in which developers engage, rather than the volume, correlates positively with the innovativeness of their future code. Notably, exposure through weak interactions (e.g., starring) emerges as a stronger predictor of future novelty than via strong ones (e.g., committing).

CCS Concepts: • **Software and its engineering → Open source model**.

Additional Key Words and Phrases: Open Source Software, Software Creativity, Social Network Analysis

ACM Reference Format:
Hongbo Fang, Patrick Park, James Evans, James Herbsleb, and Bogdan Vasilescu. 2025. The Strength of Weak Ties among Open-Source Developers: In the Python Ecosystem, Diversity via GitHub Stars Predicts Innovation Better Than Diversity via Commits. *ACM Trans. Softw. Eng. Methodol.* XX, XX, Article XX (XX 2025), 42 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Think of examples of big software innovations. You might name the Netscape browser (created the visual web and fueled mass internet adoption), Hadoop MapReduce (enabled "big data" processing), Photoshop (made digital image manipulation mainstream), Netflix (killed video rental stores), Uber ("revolutionized" transportation), or perhaps more than anything else, the World Wide Web (the global information network that connected humanity and fundamentally changed how we access knowledge).

Authors' Contact Information: Hongbo Fang, hongbofang@uchicago.edu, Carnegie Mellon University, University of Chicago, USA; Patrick Park, patpark@cmu.edu, Carnegie Mellon University, USA; James Evans, jevans@uchicago.edu, The University of Chicago, Santa Fe Institute, USA; James Herbsleb, jim.herbsleb@cmu.edu, Carnegie Mellon University, USA; Bogdan Vasilescu, vasilescu@cmu.edu, Carnegie Mellon University, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

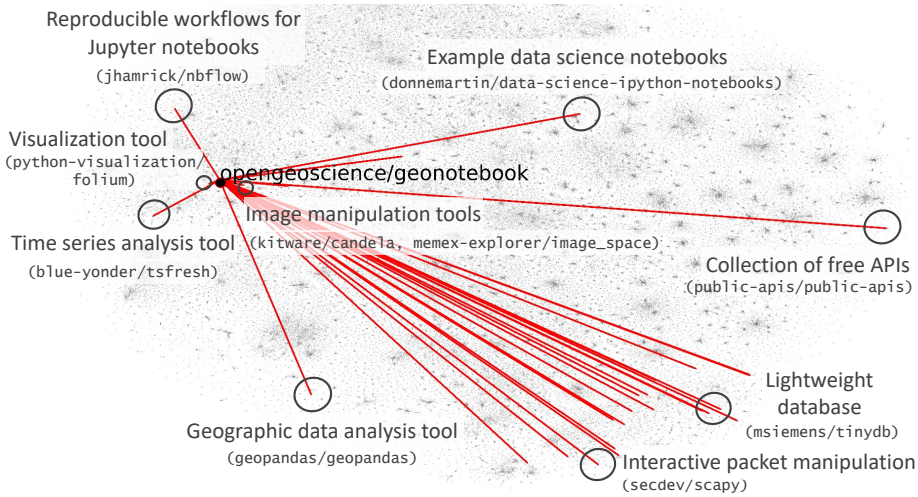


Fig. 1. t-SNE visualization of the embedding space for weak ties (details in Section 4.4), depicting all the weak ties of the GeoNotebook Python project in our sample. We highlight some that seem influential for the design of the focal project. All repositories in the graph can be accessed via *https://github.com/(repo_owner/repo_name)*, where *repo_owner/repo_name* should be replaced with actual repo_slug. (e.g., geopandas/geopandas)

Besides their undeniable impact, what all these innovations have in common is, perhaps surprisingly, that *none was highly novel in a pure sense*. Instead, they succeeded by recognizing latent potential in existing components that others had overlooked or dismissed as insufficient. This reveals a fundamental paradox in software innovation: breakthroughs can emerge not only from inventing entirely new concepts, but from developing the insight to see how familiar pieces can fit together in unexpected ways.

As he remarkably candidly notes in his book "Weaving the Web" [7], Berners-Lee didn't invent hypertext, markup languages, or networking protocols, but he saw how HTTP, HTML, and URLs could create a self-reinforcing web that previous hypertext systems had failed to achieve. The web browsers didn't invent hypertext or networking either, but they made the connection between them obvious in retrospect. Photoshop didn't invent image processing algorithms or digital manipulation, but unified traditional darkroom techniques with mathematical transforms in an intuitive interface. Uber didn't invent transportation technology, but cleverly orchestrated GPS, mobile payments, and dispatch algorithms that had existed separately for years. Finally, Netflix didn't invent video streaming or recommendation systems – they combined content delivery networks with collaborative filtering at a moment when bandwidth made it viable.

What distinguishes many impactful software innovations, then, isn't pure novelty but rather a form of *combinational creativity* [9] – the ability to perceive new relationships between existing elements and to pursue combinations that may seem counterintuitive. Many innovators aren't necessarily the first to create the individual components, but they are the first to see the gestalt that emerges from their combination. This suggests that thinking of breakthrough innovations as pure inventions may actually mislead us about where a lot of transformational software comes from. Perhaps a highly valuable skill isn't creating something from nothing, but developing the pattern recognition to spot which existing pieces, when properly orchestrated, can produce emergent behaviors that transcend their individual capabilities. Much real innovation, thus, lies not in the novelty of the parts, but in the non-obvious wisdom of their assembly. In fact, this

process of innovation through novel recombination is not specific to software – it appears to be a fundamental pattern in virtually all domains of human innovation, and it has been theorized and studied empirically in many contexts, including business, medicine, science, and technology [47, 79, 82, 94].

But while it's clear that innovation drives the software industry [33], and that great software engineers (among many other attributes) should be "*[able to] generate novel and innovative solutions based on the context and its limitations*" [60], it's still unclear how to develop one's ability to combine existing things in clever ways. Such combinational creativity "*typically requires a very rich store of knowledge, of many different kinds, and the ability to form links of many different types*" [9]. How to develop these? Although empirical evidence is scarce [44, 70], one might theorize that mixing knowledge depth with knowledge breadth is the key [24, 93], since these characteristics make software professionals more successful in general [30, 86]. According to this thinking, deep engagement with the software development process, ideally through "deliberate practice" [5], should help build expertise, while broad exploration in many different contexts should help build vocabulary.

Our paper goes one step further in this line of thinking, revealing a stronger and more nuanced effect of breadth than one might expect. Across over 37,000 Python projects on GitHub, while we confirm an association between knowledge breadth and software innovation, we find that **exposure through minimal-effort interactions like starring repositories emerges as a stronger predictor of future innovation than high-effort engagements like writing code**. This finding is even more striking given that most developers report starring repositories as a way of showing appreciation [11], not as a form of learning. But our result suggests that this casual, passive observation of diverse projects – essentially *digital lurking* – is more strongly associated with developers' future project innovativeness than direct, intensive engagement and active learning.

Much like the unexpected advantage of weak ties in job searches, which challenged conventional wisdom of social networks at the time [42], this "strength of weak ties on GitHub" represents a paradox worth exploring: Sometimes, developers' least-effort interactions with projects may contribute the most to their innovation capacity. Our research investigates this surprising mechanism, examining how the structural characteristics of developers' interaction networks correlate with the innovativeness of the software artifacts they subsequently create.

Below, we draw from established theory to formulate our hypotheses (§2); propose a novel software innovativeness measure, reconstruct strong- and weak-tie interaction networks (via commits, issues, and starring) for a project's core team, and estimate the amount and diversity of knowledge accessible to the core developers via these networks (§4); build regression models to test the association between network structural properties and software innovativeness (§5); and discuss the implications of our results (§6).

2 Theory and Hypotheses

In a nutshell, developers on the GitHub platform engage in a wide range of interactions, corresponding to ties of varying strength in a network sense, that create opportunities for knowledge acquisition via social learning. This knowledge can influence the innovativeness of the solutions they create. Before diving into the details of the mechanism, let us clarify some key constructs.

Innovation as Novel Recombination. As discussed above, we view software innovation as often emerging from *novel recombinations* of existing components, libraries, and patterns, following the long-standing view that "[business] innovation combines factors in a new way, or that it consists in carrying out new combinations" [79], which has fueled

much research in the social sciences (e.g., [47, 82, 94]). As a result, the line between *novelty* and *innovation* becomes fuzzy, thus we will use the two terms interchangeably, in contrast to some prior work that distinguishes them [6, 22].

Creativity as Antecedent of Innovation. How to facilitate the emergence of innovation is unclear, although a common argument centers on enabling creativity as a precursor. *Creativity* and *innovation* are closely related but distinct constructs: creativity is the ability to generate novel ideas, while innovation transforms novel concepts into tangible outcomes. It is expected that a software team that is more creative has access to richer knowledge (e.g., knowledge about the problem and technical approaches to solving it), and is better able to combine pieces of that knowledge to create an innovative solution [32]. Of course, there are many facets of creativity, including as an individual personality trait [54]. In this paper we focus only on one. Specifically, we argue that creativity at the team level is a function of the knowledge networks of the team members and, while we cannot measure creativity directly, we can expect to see differences in the teams' creative outputs (i.e., more novel software), that are associated with differences in the structure of those knowledge networks.

Knowledge Acquisition via Social Learning. Historically, and especially with the advent of "social coding" platforms like GitHub, participation in open source has been rife with opportunities for social learning [23], i.e., the process by which individuals acquire new knowledge, behaviors, skills, or attitudes through observation, imitation, and interaction with others. Learning, including by observing what others are doing [105], has been [57, 58] and remains [41] among the most important motivations for people to contribute to open source. In addition, modern code hosting platforms offer many opportunities for users to interact (e.g., collaboration on a shared codebase, issue discussions, code review). Through social media-like functionality (e.g., "following" [8]) and many available signals that offer a high degree of transparency [25] (e.g., repository badges [91]), the platforms also facilitate users quietly observing and being influenced by others' behaviors.

These interactions with other individuals [97] and with the artifacts they create [18] serve as channels for information dissemination and knowledge exchange. Indeed, prior studies have documented many such social learning effects, including choosing an open-source license [73, 84], discovering and adopting emerging tools [59, 83], and learning new software design principles and programming skills by reviewing code authored by others [3, 71].

Therefore, we expect that an increase in the volume of interactions among developers should correspond to a higher amount of knowledge transfer, thus better preparing developers for future innovations. Thus, we hypothesize that:

$\mathbf{H}_1$. *The more interactions developers have with other developers and projects, the more innovative their projects are.*

Still, if innovation requires an increased "vocabulary" of knowledge bits accessible for recombination, even high volumes of interaction might not be sufficient if the corresponding information is redundant. This can happen, e.g., when one works primarily within a narrow domain. Instead, we expect that accessing diverse knowledge acts as a catalyst for innovation, enabling individuals to integrate disparate concepts and to develop unconventional and novel products [80]. Outside of software engineering, Tortoriello et al. [90] observed a positive association between employees' access to diverse knowledge within the research departments of high-tech companies and their innovation levels. Similarly, Abdul Basit and Medase [1] found that the integration of knowledge from both internal research teams and customers in the public sector enhances innovation. Thus, we hypothesize that:

$\mathbf{H}_2$. *The greater the informational diversity of developers' past interactions, the more innovative their projects are.*

The Strength of Weak Ties. In a social network, ties can have varying strength, reflective of real-world factors like the duration of shared interactions, emotional depth, level of intimacy, and amount of reciprocal exchanges [42]. In that

sense, "weak" ties involve infrequent and less intimate interactions, while "strong" ties are characterized by higher frequency of interaction and intimacy.

In the late 1960s, sociologist Mark Granovetter uncovered a counterintuitive finding that would revolutionize our understanding of social networks: When searching for jobs, people relied more on casual acquaintances (weak ties) than close friends or family (strong ties). His groundbreaking theory on the "strength of weak ties" [42] revealed that these seemingly superficial connections often serve as critical bridges between different social circles, providing access to novel information unavailable within one's immediate network. This pattern arises because individuals within the same network neighborhood have numerous opportunities for interaction and often share many mutual ties. This shared social context leads to increased similarity in behaviors, interests, and other characteristics. Consequently, the information disseminated via strong ties risks becoming less novel and less valuable (the discussion of echo chambers on social media [56] is a prime example of this degradation of information quality). In contrast, information acquired via weak ties, or cross-group connections, tends to originate more from individuals with diverse backgrounds and knowledge bases, making it more likely to be novel and of greater value [42]. This mechanism has been extensively validated in social contexts, including generation of innovative ideas [14], the diffusion of news feed content [4], and job seeking in the digital age [75].

Similarly, we expect that the software-related knowledge transferred through weak ties is more valuable for the creation of innovative software projects. Thus, we hypothesize that:

H$_3$. *The more the informational diversity of developers' past interactions is due to weak ties, the more innovative their projects are.*

3 Related Works

3.1 Innovation in open-source software

Prevailing approaches to measuring innovation predominantly rely on human evaluation or criteria challenging to operationalize, thereby impeding the assessment of software innovation on large scales. Consequently, quantitative methods for evaluating software innovation remain relatively scarce and often exhibit coarse granularity. For instance, Dong et al. conducted an empirical investigation into the pace of innovation within open-source software, utilizing the number of software releases as a proxy for innovation speed without differentiating between individual releases [31]. More recently, Fang and colleagues introduced a novel approach, utilizing the unique combinations of software packages as an indicator of software novelty [36]. This method offers a scalable computational measure of novelty, supported by a robust theoretical foundation within the innovation literature. In this paper, we adopt this approach to measuring software innovation.

3.2 Social Ties, Knowledge Flow and Innovation

Participation in open-source software (OSS) development constitutes a social learning process [23], wherein social ties among OSS developers serve as channels for information dissemination and knowledge exchange. Developers acquire knowledge about emerging tools and projects by monitoring trends in the OSS community [83]. Furthermore, they gain insights into software design principles and programming skills through the review of code authored by peers [3, 71]. Interactions with project maintainers also facilitate learning about software management practices and specific implementations [17].

The knowledge accessible to developers influences their decisions and practices in software development. For example, Singh et al. and Peng et al. found that developers are more likely to choose an open-source license adopted by projects to which they are socially connected [73, 84]. Similarly, Lamba et al. revealed that developers are more likely to adopt an automated tool if they are connected to other developers who already use it, an effect observed for both collaboration ties and project-watching connections [59].

Accessing *diverse* knowledge is particularly important for innovation, as it facilitates the integration of ideas from distant domains that would otherwise be unlikely to be combined, thereby increasing the potential for generating previously unexplored ideas [94]. Empirical evidence suggests that access to diverse knowledge enhances innovative output in both industrial settings [80, 90] and the public sector [1]. Our work extends this line of inquiry to OSS, and specifically examines the role of tie *strength* in moderating the diversity-innovation relationship, a dimension that has remained largely unexplored in the software engineering context.

3.3 Social Network Analysis in Open Source Software

Social network analysis (SNA) has become a prominent methodology employed by researchers in the field of open-source software (OSS) to study social relationships. Early contributions by Crowston and Howison investigated the diverse structures of developer communication networks within bug tracking systems, revealing significant variations in communication networks in terms of activity centrality and interaction modularity [23]. Subsequent studies have further examined the evolution of communication network structures in OSS projects [62] and investigated how different network structures correlate with the effectiveness of bug fixes and overall project success [63, 102].

Another line of work examines collaboration networks among developers, considering them as conduits for knowledge exchange and resource sharing. Research in this area investigates the association between a project's or developer's network position and their success. Grewal et al. found that the number of commits received by a project is associated with both the network position of the project and its managers [43]. Similarly, Fershtman and Gandal identified a relationship between a project's closeness centrality in the project-by-project network and its likelihood of being selected as the "project of the month" [38].

Tan, Singh, and their colleagues adopted social capital theory to empirically evaluate multiple network centrality metrics and their associations with a project's popularity and productivity [85, 89]. Furthermore, networks can also elucidate developer activities. For example, Qiu et al. found that social networks influence the sustained engagement of OSS developers, with effects varying among individuals of different genders [74].

The study of developer social networks extends beyond those formed on code hosting platforms and is not limited to direct interaction. For instance, Yang et al. examined the developer follower network on the Ohloh platform and investigated its association with project success [104]. Additionally, Hu et al. analyzed developers' star, fork, and follow activities to quantify their influence [49].

While this body of work has established that network position matters for project success metrics such as popularity and productivity, it has largely not examined whether and how tie *strength* shapes the *diversity* of knowledge accessible to developers, nor how this in turn affects the innovativeness of the projects they create. Our work addresses this gap.

4 Methods

Next we give a high-level overview of our approach, before diving into operationalization details.

4.1 Overview / Intuition

Network Construction. We theorized above that (1) the interactions developers have with each other and with each other's artifacts, i.e., the ties they form, create opportunities for knowledge transfer and facilitate social learning; and (2) "strong" and "weak" ties may play different roles in this knowledge diffusion process. To operationalize these concepts, we first construct three interaction networks where every node is a project (i.e., GitHub repository), and the links between two nodes encode different interactions the developers of one project had with the other project.

There are many ways in which two open-source developers can interact, both in and outside the GitHub platform. Clearly, it's not possible to capture all interactions at scale, as projects may use, e.g., a diversity of communication channels, including private ones. Instead, we consider three representative examples of interactions that (a) vary substantially in effort, thus can be expected to reflect varying levels of knowledge flow from the target project to the author of the action (i.e., to encode person–to–project ties of varying strength, which we later project to construct our project–to–project networks), (b) are core features of the GitHub platform, thus are commonly used: making *commits* to a codebase (direct pushes or merged pull requests), posting or participating in *issue discussions*, and *starring* repositories.

Among the three, commits typically require the most effort and starring the least. This ranking should correspond to the depth of the knowledge that may transfer as a result of each action. Intuitively, while commits vary in size and content [2], an "average" commit should require a significant understanding of the project's technical details [97], indicating a path for considerable knowledge flow between the project and the commit author. Issue threads primarily discuss feature suggestions, bug reports, and user support [66]. Thus, while engaging in issue discussions can reflect a deep understanding of the project, on average we expect that it requires (and reflects) a less deep understanding of the project compared to making changes to its codebase. Finally, starring a project is often done as a sign of appreciation, a bookmarking attempt, or an intent for later use [11]. Starring a repository indicates at least some awareness of the project, but on average probably much more basic understanding than the other two.

Note that we consider only the actions of core developers in a focal project, identified heuristically as those contributors who authored at least five percent of all commits, with a minimum of 10 commits in total. Our operationalization is validated with alternative thresholds for core developer identification in the appendix (See Section G). Conceptually, in open source, core developers wield the most influence over a project's technical decisions and development trajectory [76]. Empirically, in our sample core developers also authored the vast majority (83%, based on a random sample[1] of 100 commits that import new packages) of commits importing new packages into projects (which we use to compute innovativeness, as described below). Thus, we expect that the knowledge they had access to prior to working on a focal project (as opposed to peripheral contributors) has the most influence on the design and innovativeness of that project.

Network Measures. Using our network data, we compute variables related to the network position of each project to represent the *amount* and *diversity of knowledge* that may be accessible to its core developers through each network. For example, consider the commit-based project–to–project network. A node (project) in this network may have many outgoing connections, indicating that its core developers have collectively committed to many other projects prior to joining the focal project. That is, they have first-hand experience with the technologies used in those other projects, and they can probably draw from that experience (knowledge) now to influence the development of the focal project. Similarly, we can reason about the other two networks constructed from issue discussions and starring repositories.

How to measure the diversity of knowledge is less obvious. Conceptually, the amount and diversity of knowledge are related yet distinct. For example, a project whose core developers are connected to a large number of other projects

[1]This random sample is available in the replication package

within the network *may* have access to diverse knowledge. However, the actual diversity of this knowledge remains contingent on the nature of these connections – if all linked projects contain similar knowledge, diversity may be limited; e.g., if they're all related to visualization, that doesn't say much about experience with TENSORFLOW.

To capture this diversity, we adopt an approach analogous to word embeddings, wherein semantic similarity between words is inferred from their vector representations [68]. The key idea is that we compute project graph embeddings and use embedding distance to quantify the similarity of gainable knowledge among projects. The intuition is that nodes that frequently appear together in the same random walk while learning the embeddings represent projects that are often contributed to or interacted with by the same developers, i.e., are related in some informational sense. This could be because they share similar functionality, are used in similar contexts, or are part of the same development ecosystem.

Dimensionality Reduction. Thus far we have been treating our three networks (commits, issues, and stars) as separate. However, one can expect that the variables we compute based on these networks are correlated to some extent. Moreover, one can imagine considering other types of interactions (i.e., networks) besides our three, which would result in yet more variables. To prevent issues of multicollinearity in our subsequent regressions, which can both complicate the interpretation of the estimated coefficients and reduce statistical power [26, 96], we use Principal Component Analysis (PCA) [101] to decompose the variables representing the *amount* and *diversity of knowledge* into orthogonal components. As we show below, this produces a clear decomposition of our original variables into two components capturing strong and weak ties, which we use in our models instead of the original variables.

Project Innovativeness Measure. Fang et al. [36] introduced a measure of a software project's novelty, or innovativeness, as a function of the combinations of packages the project imports. For example, while certain libraries, such as NUMPY and TENSORFLOW, are frequently used together, others, like NUMPY and REQUESTS, are less commonly combined. The measure estimates how atypical combining two packages is, and by extension how atypical is the overall set of packages being used as dependencies in a project, by comparison to what could be expected by random chance. Projects that import more atypically paired packages are deemed more innovative as they reuse packages in nontraditional ways.

We maintain this framing of **innovation through novel recombination**, and similarly consider **packages as the unit of recombination**. Packages are designed to be reused; they're typically substantial enough to provide real functionality while remaining composable (individual functions are too small – they lack sufficient context and capability to be meaningfully recombined; entire applications are too large); and they're usually semantically coherent, i.e., they tend to be organized around conceptual domains (e.g., REQUESTS encapsulates the entire complexity of HTTP communication in a reusable form, while PANDAS bundles decades of data manipulation knowledge into a coherent interface). Alternatively, one could consider something like "concepts" as the unit of recombination, as in some prior science-of-science work [47]. However, it wasn't clear how to automatically extract concepts from software, so we chose packages as the unit instead.

Our measure of project innovativeness, while similar in spirit to Fang et al.'s [36], is based on learning embeddings of packages and addresses two important limitations of the original measure. First, the original measure only captures pairwise combinations between packages (and aggregates up to the project level from that), neglecting the possible interactions among related packages. However, packages often form "stacks" comprising more than two libraries that complement each other and are often reused together. Our embeddings-based approach considers more context and is expected to better capture a package's technical use, function, and "stack." Second, computing the original measure

requires random reshuffling of the global dependency network, which is computationally prohibitive at the scale of our study. In contrast, we compute the cosine similarity of embeddings, which is faster.

Regression Analysis. Putting everything we discussed so far together, our analysis involves modeling the variation in project innovativeness across our sample as a function of the variables we discussed above, controlling for known covariates. While not causal,[2] this analysis will still allow us to test the extent to which the observational data is consistent with the causal paths theorized above (about determinants of combinatorial creativity).

In the remainder of this section we give lower-level operationalization details for the steps above, including validation checks.

4.2 Sample Selection

We combine data from World of Code (WoC) [64] and the GitHub API. WoC is arguably the most comprehensive record of open-source projects and their commit history, and it also contains historical import-based dependency information for projects in the most popular programming languages, which we use when computing project innovativeness. To keep our analysis tractable, we focus only on Python projects, i.e., the project contains more than 10 Python source code scripts. We chose Python due to its immense popularity [88] and its widespread use across various domains by developers from diverse backgrounds [77]. While not allowing us to directly generalize our findings, the considerable variability in the Python ecosystem across many characteristics should ensure points of overlap with many other ecosystems.

Because many public repositories serve primarily as code archives rather than active software development projects, and to address the limitations and complexities associated with GitHub data [53], we excluded forked repositories and repositories containing fewer than ten commits. In addition, repositories whose first commit occurred before 2008 (i.e., prior to the launch of GitHub) or after 2021 were removed from the sample. As the primary data collection for this study was completed in 2022, the latter restriction ensured that each repository had at least one year of observable history in WoC, which was necessary for inferring project innovativeness. After applying these selection criteria, the final sample comprised 1,048,733 projects.

Next, we removed projects that imported no more than one package prior to the end of the observation period. Consistent with the rationale of Fang et al. [36], Python standard libraries were excluded from the analysis. Following this filtering step, the sample size decreased further to 727,979 projects.

We additionally excluded projects for which no core developer could be identified. Applying our primary and more stringent definition of core developers: contributors responsible for at least ten commits and at least five percent of a project's total commits, as described in Section 4.1, reduced the sample size substantially to 610,149 projects. This filtering criterion had no effect on the sample size under the alternative definition of core developers presented in Section G, in which our main results are validated.

For the remaining projects, we queried the GitHub API to collect historical information on commits, issues, and stars, as the latter two data types are not available through WoC. For each event, we recorded the associated author and timestamp. These data were subsequently used for network construction, as detailed in Section 4.3.

Finally, we removed projects exhibiting outlier values on either dependent or independent variables. The resulting sample used in Model I of Table 3 comprised 589,999 projects, representing a reduction of 3.3% from the previous sample of 610,149 projects. For Models II–IV, which incorporate diversity measures, we additionally excluded projects whose core developers had interacted with no more than one project within each interaction modality considered (i.e.,

[2]Testing our hypotheses experimentally is infeasible, and identification is otherwise unclear from observational data like ours, since there is no clear "intervention."

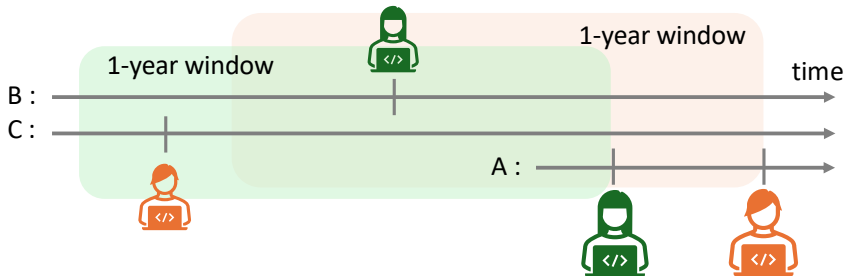


Fig. 2. Two core developers Green and Orange started contributing to a focal project A on different dates. We record an edge from A to B (i.e., B could be a source of knowledge for A), because Green interacted with project B in the previous year. We don't record an edge from A to C, because Orange interacted with C too far into the past.

commits, issues, and stars; we explain this rationale in Section A). Consequently, the sample size decreased from 589,999 to 167,879 projects after excluding projects with insufficient interactions in the commit network, further decreased to 55,376 projects after filtering based on the issue network, and finally reached 37,451 projects after applying the star network filter.

In addition, we heuristically identified and filtered out bot accounts [29, 100], which would distort our networks and subsequent measures, as they tend to have a lot of activity. To this end, we searched for keywords in user logins, such as *-bot* or *-robot*, and labeled those users as bots after manual review. Additionally, we manually inspected the profile descriptions of the top 100 most active user accounts in our sample, ordered by the number of commits, to ensure the removal of highly active bot accounts that may have escaped the naming convention heuristics.

As a validation step, we randomly sampled 100 projects from the analyzed dataset and manually examined whether the identified core developers were bot accounts. Among the 189 identified core developers, only one account (0.53%) was determined to be a bot, affecting only a single project (1%)[3]. These results suggest that our bot-filtering procedure was sufficiently effective and that residual bot activity is unlikely to meaningfully bias the study's estimates.

4.3 Network Construction

Each of the three actions we consider (making commits, participating in issue discussions, and starring) can be thought of as a person–to–project tie. Aggregating across all developers and projects in our sample leads to three person–to–project bipartite networks, one for each action type. However, since our analysis is at the project level, we construct three directed project–to–project networks by way of projection, to represent possible flow of knowledge across projects. As discussed previously, we consider only actions initiated by core developers (e.g., which other repositories they committed to or starred in the past). In addition, we restrict the scope to actions in the recent past (e.g., within the last year, we validate with alternative period length in the appendix), as we expect older actions to be less relevant. This time window is developer-dependent. That is, we define an edge $A \rightarrow B$ if and only if a core developer of project A interacted with project B within one year prior to their initial commit to project A , as illustrated in Figure 2. For robustness, we also validate our results using alternative temporal windows of 6 months and 24 months (See Section G). The weight assigned to each edge corresponds to the number of core developers from project A who engaged with project B via a specific type of interaction.

[3]This random sample is available in the replication package

We also validate the relative strength of the three networks using their transitivity values.[4] We adopt this structural metric because the original concept of weak ties [42] characterizes network strength through structural properties and derives its implications from network topology. Across the three networks, we observe an approximately one order of magnitude (10×) difference in their transitivity values. Specifically, the commit network displays the highest levels of transitivity, followed by the issue network, while the star network exhibits the lowest level. This relative difference remains consistent after controlling for differences in network density across the three networks (See Section B).

4.4 Network Measures

Next, we compute variables related to the network position of each project to represent the *amount* and *diversity of knowledge* that may be accessible to its core developers through each network.

First, we use the number of connected projects as a proxy for the **amount of knowledge accessible** to a given project. More formally, for every project, we compute its out-degree centrality within each of the three networks, defined as the total sum of the weights of all its out-edges. This step resulted in computing the three variables listed in Table 2 under ***Degree Variables (original)***.

Second, we use the Node2Vec graph embedding algorithm [45] to generate vector representations (embeddings) for each node in the network based on their topological position. The process begins by generating sequences of random nodes from the graph using random walks. Starting from a random node in the network, we sample the next node by randomly selecting a direct neighbor. The likelihood of selecting each node is proportional to the edge weight between the candidate next node and the current node. This process continues until we have selected enough nodes to complete the walk (in our study, each walk comprises 20 nodes). We repeat this process to generate multiple walks, or sequences of nodes (in our study, we generate 30 walks starting from each node, and set p = q = 1 to balance exploration and exploitation, and breadth-first and depth-first search). Subsequently, we use a skip-gram model [67], commonly used to generate embeddings for words in natural language, on the generated walks to learn embeddings for each node. In our study, we set embedding dimension to be 512, window size to be 3 on each side, min_count being 1 to keep all appearing nodes, and train the model for 20 epochs.

Next, for each project in the network, we examine all other projects that receive a directed edge from the focal project (denoted as set P). Our **knowledge diversity index**[5] is the average pairwise distance between any two projects in P: $D = \frac{\sum_{i,j \in P, i \neq j} -2 * sim(v_i, v_j)}{|P| * (|P|-1)}$, where i and j are projects in P, and v_i and v_j are their vector representations from the Node2Vec model. The distance between v_i and v_j is the negation of the cosine similarity of the two vectors. This step resulted in computing the three variables listed in Table 2 under ***Diversity Variables (original)***.

4.5 Dimensionality Reduction

Since they are designed to capture different concepts, we run PCA separately for the three degree variables and the three diversity variables. As standard, we first log-transformed all input variables and scaled them to a mean of zero and a standard deviation of one.

The PCA resulted in three components for the degree variables and another three for the diversity variables, with the proportion of variance explained by each component listed in Table 1 (top). Inspecting the table, we observe that the first two PCs cumulatively explain over 80% of the variance in both groups, thus we decided to retain only two components

[4] In network science, *transitivity* measures the ratio of triangles to triads in a network.
[5] The index is undefined for projects with out-degree centrality below two. We exclude these from our regression when considering diversity metrics as independent variables.

Table 1. *Top*: Proportion of variance explained by each principal component. *Bottom*: Loadings of the principal components onto the original degree and diversity variables.

	Out-deg. centrality			**Diversity index**		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
Variance Explained	0.64	0.22	0.14	0.51	0.29	0.19
Cumulative Variance	0.64	0.86	1.00	0.51	0.81	1.00
D_{commit}	0.60	−0.45	0.67	0.63	−0.36	0.69
D_{issue}	0.61	−0.28	−0.74	0.65	−0.24	−0.72
D_{star}	0.52	0.85	0.11	0.43	0.90	0.08

each for the degree and diversity variables. That is, we will use these PCs (listed under ***Degree Variables (post-PCA)*** and ***Diversity Variables (post-PCA)*** in Table 2) instead of the original variables in our regression models below.

To interpret the two PCs we turn to Table 1 (bottom), which lists the loadings of the principal components onto the original variables.[6] Inspecting the table we make the following observations. First, the loadings onto the first principal component (PC1) are relatively consistent across networks for both the degree and the diversity variables. Thus, we interpret PC1 to represent the average degree (or diversity of ties) of the project across the three networks. For example, drawing on our theoretical framework (Section 2), a project for which $PC1_{\text{degree}}$ is high is expected to have many ties (on average, across the three networks), i.e., its core developers should have more sources from which to draw inspiration. Similarly, the ties of a project for which $PC1_{\text{diversity}}$ is high are expected to span more of the knowledge embedding space we used to estimate tie diversity, i.e., its core developers could have access to more varied sources to draw inspiration from.

Second, the loadings onto the second component (PC2) show a descending trend across networks, again similarly for both sets of variables. Notably, the highest loading on the degree (or diversity) comes from the star network, followed by the issue network, while the lowest comes from the commit network. Thus, we interpret PC2 to represent the strength of network ties in the degree (or diversity of ties) metric, i.e., the **strength of weak ties**. A project for which $PC2_{\text{degree}}$ is high is expected to get more of its connectivity through the star network. Analogously, relatively more of the diversity of knowledge accessible to core developers in a project for which $PC2_{\text{diversity}}$ is high can be attributed to the star network (weak ties) compared to the commit or issue networks (stronger ties).

It may seem counterintuitive to reason about PC1 and PC2 jointly, especially in a regression modeling framework. What does it mean to vary PC2 from low to high, for instance, while holding PC1 fixed? Can PC1 and PC2 even vary independently? They can in theory, because they are orthogonal by construction, but are the combinations of low-high values of PC1 and PC2 observable in the real-world data? After all, regression is an interpolation mechanism. We illustrate this space of interpretations of PC1 and PC2 for the diversity variables with the artificial examples in Figure 3 (see Supplementary for real-world examples in each quadrant). In the top left quadrant (PC1 and PC2 both high), the project core developers can tap into a diverse knowledge space via either strong or weak ties. In the top right (PC1 high and PC2 low), the overall diversity of information accessible is just as high, but it's the strong ties that are the source of it. In the bottom left (PC1 low and PC2 high), the overall diversity of information available is lower, but whatever diversity there is, it can be attributed mostly to the weak ties. Finally, in the bottom right (PC1 and PC2 both low), weak ties contribute little to an otherwise low diversity of information.

[6] Principal components are linear combinations of the original variables, with the loadings indicating the contributing coefficients. For example, the first PC for the degree variables, $PC1_{\text{degree}}$, uses the coefficients in the first column in Table 1 (bottom): $PC1_{\text{degree}} = 0.60 * Deg_{\text{commit}} + 0.61 * Deg_{\text{issue}} + 0.52 * Deg_{\text{star}}$.

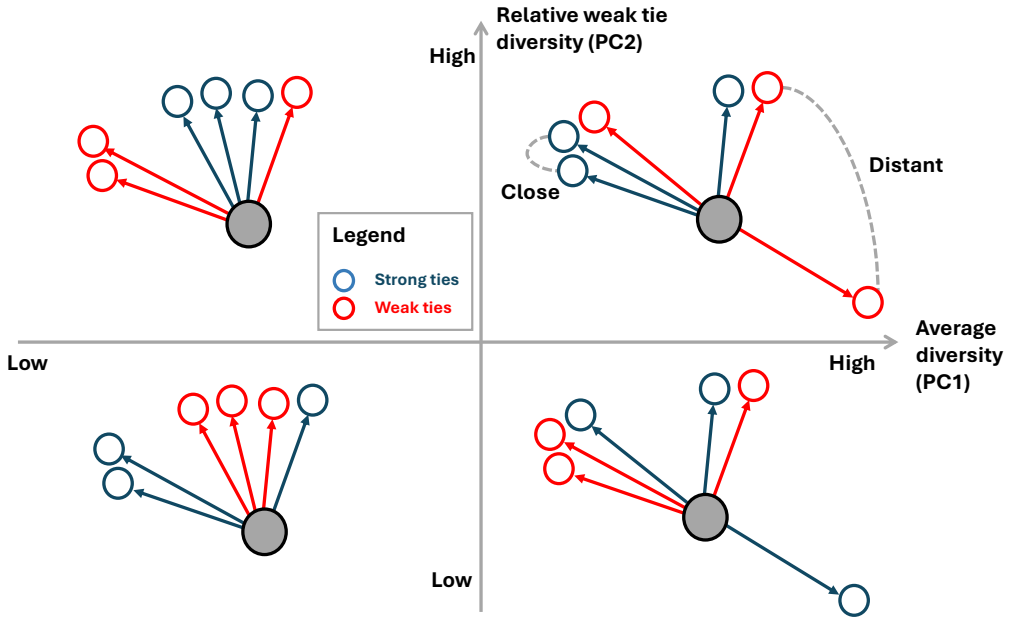


Fig. 3. Illustration of the four possible combinations of low-high values for $PC1_{\text{diversity}}$ (corresponding to average tie diversity) and $PC2_{\text{diversity}}$ (corresponding to the relative diversity attributable to weak ties).

Going back to our hypotheses (Section 2), we expect the left quadrants, where weak ties are strong, to be most associated with innovativeness. As an anecdote, consider the example in Figure 1 on the first page. The GeoNotebook project is a Jupyter notebook-based environment for interactive visualization and analysis of geographic data. Interestingly, GeoNotebook's core developers have previously starred (without otherwise contributing to) many GitHub projects spanning a variety of seemingly related topics, including a collection of free APIs, example data science notebooks, reproducible workflows for Jupyter notebooks, visualization tools, time series analysis tools, geographic data analysis tools, and infrastructure tools for interactive packet manipulation. One can expect that this diverse space of ideas provided some inspiration, knowingly or unknowingly, for the design and implementation of GeoNotebook. Our analysis below tests to what extent there is evidence supporting this mechanism at scale, beyond this anecdote.

4.6 Project Innovativeness Measure

For each project, we extract the sequence of packages imported[7] (precomputed in World of Code) and use a skip-gram model to generate positional embeddings for each package based on these dependency relationships. These embeddings place the package in an optimal location relative to all other packages imported in the same project, based on their co-import relationships. Unlike the random walks used in the Node2Vec model to generate embeddings for the three interaction networks, there is no inherent order in the sequence of packages imported within the same project. Therefore, we set the window size in the skip-gram model to be sufficiently large (we set the window size to be 25 on each side, and

[7]We ignore built-in Python libraries like Fang et al. [36] did, and consider only packages published in the PyPI registry, using the explicit links back to GitHub recorded there.

Table 2. The definitions of the variables in our models.

Outcome Variables	
Innovativeness	The extent to which a project imports packages that were typically not combined in the past, computed based on all the packages imported until the end of 2021.
Degree Variables (original)	
Deg_{Commit}	The sum of edge weights from the focal project to the others in the commit network.
Deg_{Issue}	The sum of edge weights from the focal project to the others in the issue network.
Deg_{Star}	The sum of edge weights from the focal project to the others in the star network.
Degree Variables (post-PCA)	
Deg_{Ave} ($\mathbf{H}_1$)	The first principal component resulting from the PCA analysis over three original degree variables, representing average network degree across the three networks.
Deg_{Weakness}	The second principal component resulting from the PCA analysis over three original degree variables, representing the degree from weak ties.
Diversity Variables (original)	
Div_{Commit}	The diversity of projects to which the focal project has a directed edge in the commit network.
Div_{Issue}	The diversity of projects to which the focal project has a directed edge in the issue network.
Div_{Star}	The diversity of projects to which the focal project has a directed edge in the star network.
Diversity Variables (post-PCA)	
Div_{Ave} ($\mathbf{H}_2$)	The first principal component resulting from the PCA analysis over three original diversity variables, representing average diversity of connections across the three networks.
Div_{Weakness} ($\mathbf{H}_3$)	The second principal component resulting from the PCA analysis over three original diversity variables, representing the diversity from weak ties.
Control Variables	
$Year_{\text{creation}}$	A fixed effect variable indicating the year in which the project received its first commit.
Org_owned	A binary variable indicating whether the project was owned by an organizational account or not.
$N_{\text{Owner_Stars}}$	The total number of stars that the project owner received on all other repositories they owned before the creation of the focal project.
$N_{\text{Core_Devs}}$	The total number of core developers in the project, as computed before the end of 2021.
N_{Packages}	The number of software packages that the project imported before the end of 2021.

projects with more than 50 non-standard, non-customized packages imported as excluded for embedding training; each embedding has 256 dimensions and trained for 30 epochs), enabling it to consider all packages in the same sequence as its context when generating embeddings. Similar approaches to compute embeddings of software packages have been used before in software engineering research [28, 35].

To estimate the atypicality of a combination of two packages, we compute the negation of their embeddings' cosine similarity. Then, our **measure of project innovativeness** is the average pairwise atypicality of all pairs of packages imported by a project.

Validation. Although measuring project novelty in terms of the atypicality of package combinations therein has solid theoretical foundation (Section 2) in addition to application in past empirical studies of open-source software [36], this likely remains the most controversial part of our methodology. Therefore, we perform two validation checks, comparing to an existing dataset of "awesome" projects and to the previous measure by Fang et al. [36].

"Awesome" lists are community-curated lists of software. Many exist for different programming languages, platforms, application domains, etc. One can propose new entries via a pull request (PR), and typically some minimum amount of support (+1s) is required for the PR to be merged. Being innovative isn't formally a requirement for being "awesome," although as part of the same PRs submitters usually argue that the proposed entries are novel (e.g., PRs to the Python list contain a section titled "What's the difference between this Python project and similar ones?")

We scraped software mentions from a popular aggregator of "awesome" lists [87] (350k+ GitHub stars at the time of writing), and identified 760 Python projects matching our sampling criteria (§4.2), and for which we could compute atypicality scores (e.g., having at least two import-based dependencies we could identify), corresponding to about 2% of our sample. Comparing the atypicality scores between the "awesome" and remaining projects in our sample

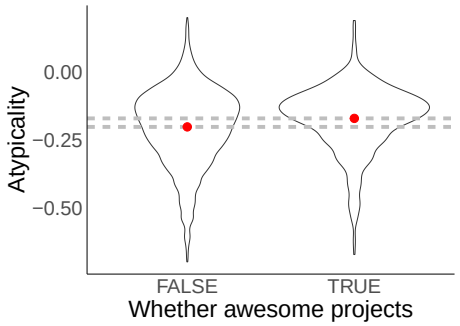


Fig. 4. "Awesome" projects have statistically significantly higher atypicality scores than the rest of our sample. The red dots represent the distribution means.

(Figure 4), we observe a statistically significant difference in means (two-sample independent t-test $p < 8.8e^{-13}$), with a small to moderate, non-trivial effect size: a project with a median atypicality score in the non-selected group will only be among the bottom 35% ordered by atypicality score among the "awesome" projects (Cohen's d = 0.22 with 95% confidence interval 0.15-0.30). Further validation suggests this effect is not confounded by variables such as project size (Supplementary). Therefore, we conclude that **our innovativeness measure reflects to some extent the developers' perception of "awesome" projects, which usually includes a dimension of novelty.**

In addition, we compute the correlation between the atypicality scores obtained using our skip-gram model and those generated using the original measure by Fang et al. [36]. Results show a moderate-to-strong [78] linear (Pearson) correlation of 0.65. Therefore, we conclude that **our innovativeness measure correlates to that of prior work to a large degree.** Furthermore, we test the robustness of our conclusions to variations in the innovativeness measure, and find that the estimated regression coefficients for the *degree* and *diversity* variables behave consistently with both measures. Thus, we only present the results obtained with our proposed measure below, but include the corresponding regression results with the original measure by Fang et al. [36] in Supplementary.

4.7 Regression Modeling Considerations

To test our three hypotheses, we use fixed-effects linear regression (one data point per project), where the response variable is our measure of *innovativeness*, and our main explanatory variables are the degree and diversity principal components discussed above: Deg_{Ave} captures $\mathbf{H}_1$ (volume of interactions the core developers had with other projects in the previous year), Div_{Ave} captures $\mathbf{H}_2$ (diversity of information accessible), and Div_{Weakness} captures $\mathbf{H}_3$ (strength of weak ties). The remaining principal component, Deg_{Weakness}, does not directly map to one of our hypotheses, but can help distinguish whether it's the volume of weak ties, or their strength, that explains more variance in our outcome measure.

We also control for several important confounding variables (Table 2): the repository owner's social standing ($N_{\text{Owner_Stars}}$), the project team size ($N_{\text{Core_Devs}}$), the complexity of the codebase in terms of number of packages imported (N_{Packages}), whether the project is owned by an organizational account (*Org_owned*), and the project age ($Year_{\text{creation}}$, modeled as a categorical variable). For model fit and diagnostics we followed standard practice, e.g., we first log-transformed the variables with skewed distributions to reduce heteroskedasticity [55]; see replication package.

Table 3. Summary of regression analysis of factors associated with differences in project innovativeness.

	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{ave} ($\mathbf{H}_1$)	0.001^{***}	$-0.001^{n.s.}$		-0.002^{**}
	(0.0002)	(0.0006)		(0.0006)
$Deg_{weakness}$	$-0.0003^{n.s.}$	$-0.001^{n.s.}$		-0.005^{***}
	(0.0002)	(0.0007)		(0.0008)
Div_{ave} ($\mathbf{H}_2$)			0.006^{***}	0.007^{***}
			(0.0006)	(0.0006)
$Div_{weakness}$ ($\mathbf{H}_3$)			0.005^{***}	0.007^{***}
			(0.0008)	(0.0008)
Controls				
Org_owned	0.027^{***}	0.017^{***}	0.018^{***}	0.018^{***}
	(0.0006)	(0.0016)	(0.0016)	(0.0016)
$N_{\text{Owner_Stars}}$ (log)	0.002^{***}	0.001^{*}	$0.001^{n.s.}$	0.001^{*}
	(0.0002)	(0.0004)	(0.0004)	(0.0004)
$N_{\text{Core_Devs}}$ (log)	0.014^{***}	0.018^{***}	0.017^{***}	0.018^{***}
	(0.0006)	(0.0016)	(0.0015)	(0.0016)
N_{Packages} (log)	0.042^{***}	0.023^{***}	0.023^{***}	0.024^{***}
	(0.0003)	(0.0011)	(0.0011)	(0.0011)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Observations	589,999	37,451	37,451	37,451
Adjusted R^2	0.054	0.059	0.063	0.064

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant

5 Results

Next we present the main results from the regression analysis formally testing our hypotheses, summarized in Table 3, as well as a series of robustness checks.

Main Regression Results.

Model I (all projects) and Model II (projects with a minimum out-degree centrality of two across all three networks) shed light on the impact of degree variables. Model I reveals a significant positive effect from average degree centrality, whereas Model II shows a negative effect. Nevertheless, effect sizes for both models are relatively small. For instance, in Model I, transitioning a project's Deg_{ave} from the 25th percentile (-1.09) to the 75th (0.53) results in a mere 0.002 increase in project innovativeness. Considering that the 25th percentile of a project's innovativeness score is -0.33, and the 75th is -0.12 in our sample, the change in project innovativeness due to variations in average degree centrality is practically negligible. Therefore, $\mathbf{H}_1$ is not supported.

> We do not find supporting evidence that the more interactions developers have with other developers and projects, the more innovative their projects are ($\mathbf{H}_1$).

In Model III, we observe a positive association between the average diversity (Div_{ave}) of ties originating from the focal project and project innovativeness. Moreover, this effect is notably stronger – nearly seven times greater – compared to that of average degree in Models I and II. Transitioning a project from the 25th percentile (-0.89) to the 75th (0.86) in terms of Div_{ave} corresponds to an increase of approximately 0.012 in project innovativeness, representing roughly a 4% change in the distribution (e.g., a shift from the median project in terms of novelty to the 54th percentile). We conclude that $\mathbf{H}_2$ is supported.

On average, the greater the informational diversity of developers' past interactions, the more innovative their projects are ($\mathbf{H}_2$).

Additionally, we observe a significant effect for the $Div_{weakness}$ variable. This suggests that, when controlling for the average diversity across the three networks, higher diversity in the "weak ties network" (i.e., star network), or stronger weak ties, corresponds to increased project innovativeness. In practical terms, the effect size for the $Div_{weakness}$ variable is comparable, albeit somewhat smaller, to that of the Div_{ave} variable above. We interpret this as supporting evidence for $\mathbf{H}_3$.

On average, the more the informational diversity of developers' past interactions is due to weak ties, the more innovative their projects are ($\mathbf{H}_3$).

When incorporating both degree and diversity variables in the same regression model, as in Model IV, the positive effects of Div_{ave} and $Div_{weakness}$ remain significant. This suggests a consistent effect stemming from tie diversity and strength of weak ties. Nevertheless, we also note a significant negative effect for Deg_{ave} and $Deg_{weakness}$ after adjusting for project diversity. These negative effects suggest that, while holding tie diversity constant, higher out-degree centrality and more ties in the "weak network" are associated with decreased project innovativeness. This observation could be attributed to the increased constraints within a local community associated with higher out-degree, as discussed by Burt [13]. Such constraints can lead to a reduced inclination for combining packages of diverse functions within the project, consequently contributing to the reduced novelty of the project. More research is needed to further test these effects.

Robustness Checks. In addition to the regressions above, we test how robust our conclusions about the effects of Div_{ave} and $Div_{weakness}$ are to differences in operational decisions.

Does Time of Measurement Matter? "No." We segmented our dataset by project inception year and conducted regression analyses with the specification of Model III in Table 3 on projects initiated in each respective year. Figure 5a shows the resulting estimated coefficients for the two diversity variables. As illustrated in the graph, the positive relationship between the average diversity of knowledge access across the three networks (represented by the blue line) remains consistent over the years, with the average estimated coefficient being statistically significantly above zero in most years. The estimated coefficient for Div_{ave} appears to also increase in recent years, though this difference is not statistically significant.

The positive relationship between diversity from weak ties and project innovativeness is generally consistent (the red line in Figure 5a). The estimated coefficient is above zero in all ten years observed, and the statistically insignificant coefficient (i.e., the lower bound of the 95% confidence interval is below zero) is likely attributable to the limited number of projects in each annual cohort (i.e., fewer than 4,000 projects in each year before 2016).

Does Project Team Size Matter? "No."

Similarly, we partitioned projects into subsets according to the number of core developers and examined the effects of the diversity variables within each cohort. As shown in Figure 5b, the effect of *Divave* and *Divweakness* remain mostly stable across projects with different team sizes, further supporting the robustness of our findings. Notably, the effect size of both variables decrease as the number of core developers increase, and even become insignificant with

four or more core developers. One possible explanation is that larger teams already possess greater internal diversity, thereby diminishing the marginal contribution of diversity derived from weak external ties.

Does Project Ownership Matter? "No." Lastly, we re-estimated the effects of the diversity variables separately for projects owned by organizational and individual accounts. Our analysis (Figure 5c) shows that both *average diversity across networks* and *diversity from weak ties* have significantly positive effects for both categories. Furthermore, the difference in effect size between organization-owned and individual-owned projects is not statistically significant.

In conclusion, the positive associations between average knowledge diversity and project innovativeness, as well as between diversity from weak ties and innovativeness, are robust.

6 Discussion

Next we discuss our results in the broader context of the literature and highlight budding directions for future research and practice.

Summary of Main Results. Established social science theory describes the mechanisms through which social network weak ties can be particularly "strong" – they can bridge disjoint parts of the network, providing access to diverse information, which in turn can be used towards better outcomes. Our study provides the first large-scale empirical evidence (that we know of) of an association between weak-tie network structure and software innovativeness. We hypothesized that the many user-to-user and user-to-artifact interactions possible on the GitHub platform, ranging from high-effort ones like making changes to a codebase to relatively trivial ones like starring a repository, correspond to network ties of varying strength that may be associated with different informational environments. In particular, we expected weak ties to be associated with greater informational diversity, consistent with the GeoNotebook anecdote in the Introduction (Figure 1).

Using a sophisticated methodology to reconstruct interaction networks, estimate tie strength, compute network informational diversity, and estimate project innovativeness, coupled with a robust statistical analysis, we found clear evidence (albeit only correlational) supporting two of our three hypotheses: on average, it's not the amount ($\mathbf{H}_1$), but rather both the diversity of information available to a focal project's core developers ($\mathbf{H}_2$), as well as the extent to which this information diversity comes from weak ties ($\mathbf{H}_3$), that explain some of the variance in how novel (atypical) the combination of software packages imported in a focal project is.

Quantity vs. Diversity of Interactions. The null result for $\mathbf{H}_1$ deserves closer attention, as it cuts against the intuition that more interactions should lead to better outcomes. We argue this is consistent with a straightforward information-theoretic argument: if additional interactions are with the same or similar projects, they contribute little new knowledge. More interactions in a redundant neighborhood do not expand a developer's information horizon; they merely reinforce what is already known. What matters is not how much a developer interacts, but how diverse those interactions are.

This interpretation is further supported by the negative effect of degree variables observed in Model IV, when diversity is held constant. This pattern aligns with Burt's [13] notion of network constraint: developers embedded in dense, locally connected networks face social pressure to conform to local norms and practices, which can suppress the inclination to combine packages in unconventional ways. High degree without diversity, in other words, may actively work against innovation rather than merely failing to promote it.

Together, these findings suggest a reframing of how developer activity on platforms like GitHub should be interpreted. Raw interaction volume, as measured by commit counts or degree centrality, is a poor proxy for innovative potential. The informational diversity of those interactions is what matters, and this distinction has practical consequences for how we design recommendation systems, evaluate developer expertise, and study knowledge diffusion in software ecosystems.

Effect Sizes. Some readers may question whether our main variables explain *enough* variance. We think small effect sizes are exactly what one should expect here, and several considerations help explain why.

First, the weak ties mechanism is *one* pathway to innovation, not the only one; alternative mechanisms jointly compete to explain the same variance. For example, a developer's *absorptive capacity* [21] may moderate how much they benefit from diverse weak ties; *organizational and community embeddedness* may independently enable novelty regardless of personal network structure; and *intrinsic project goals* determine from the outset whether a project is even oriented toward experimentation [16].

Second, our measurement of weak ties introduces its own attenuation. Starring a repository is a noisy proxy for actual knowledge absorption: many starred repositories may be bookmarked but never revisited, so our exposure variable captures *potential* rather than *realized* knowledge transfer, mechanically pushing our estimates toward zero.

Third, small effect sizes are common in social network research on innovation. Burt's foundational work on structural holes [14], Fleming et al.'s work on collaborative invention [39], and Cattani and Ferriani's [16] work on creativity consistently report similarly modest effects. The consistency of our effects across multiple strata (years, team sizes, project ownership type), coupled with statistical significance in a noisy environment, strengthens rather than undermines confidence in the underlying mechanism. Next, we discuss the implications of our results.

The Power of Exposure through Weak Interactions.

Our finding that the diversity of projects developers star is strongly associated with the atypicality of the projects they subsequently create may initially appear surprising. The primary motivation for starring a project is generally understood as signaling appreciation [11]. As such, starring appears to represent a lightweight form of interaction that reflects exposure to a project rather than sustained engagement or direct knowledge transfer. This raises an important question: can such seemingly trivial exposure meaningfully influence developers' future creativity?

Although counterintuitive, prior research suggests that the answer is YES. Dating back to as early as the 1960s, Zajonc [106] showed that repeated exposure alone can increase individuals' affinity toward stimuli. Similar effects have since been observed across a variety of domains. In political campaigns, exposure to voting-related messages has been shown to increase individuals' likelihood of participating in elections [10]. In scientific publishing, merely exposing scientists to a paper's title and abstract in review invitation has been found to increase their likelihood of citing that paper in the future [98].

Our findings provide another evidence for the influential role of exposure within software engineering. Despite often being regarded as a trivial interaction [11], starring behavior (or digital lurking more broadly) may exert a far greater influence on developers and the open-source software community than previously assumed [52].

Platform Design.

One way to interpret the strength-of-weak-ties theory in our context is that low-effort forms of exposure on the GitHub platform (i.e., we operationalize starring repositories as a weak-tie interaction rather than active contribution) are associated with greater informational diversity.

If these observed associations reflect, at least in part, developers being exposed to ideas from repositories they starred and later incorporating some of those ideas into future projects, then platform designers might consider ways of supporting broader and more diverse patterns of discovery. For example, GitHub highlights "what the community is most excited about today" through its Trending page. One possible design direction would be to incorporate informational diversity into such recommendation mechanisms, in addition to repository popularity or star growth. For instance, embedding-based approaches like ours could potentially be used to surface repositories that broaden the diversity of projects to which users are exposed. More speculatively, such recommendations could also be personalized, i.e., highlighting repositories that increase informational diversity relative to a given user's existing interaction history.

Our results also raise questions about the GitHub culture of (implicitly) rewarding code contributions the most, since one's commit history tends to be highly visible in the platform. Much has been written about how other developers [65] and even recruiters [15] use such signals. In contrast, our findings suggest that developers who engage less visibly but maintain broader and more diverse patterns of exposure across projects may also be associated with highly novel outputs. While our study cannot establish causality, this pattern points to the possibility that low-effort interactions such as starring may capture forms of exploration and awareness that are not reflected in conventional activity metrics. This raises interesting design questions about whether and how platforms might make these forms of engagement more visible alongside traditional contribution metrics.

It's also worth thinking about what tracking and giving credit to ideas, not just code artifacts, might look like in this domain. In scientific research, authors are expected to cite prior work that influenced their thinking. In turn, these citations form networks encoding invaluable information about scientific progress and scientific collaboration [40]. What might such a system look like in software development? Anecdotally, it seems like developers already use many ad-hoc ways to credit ideas, e.g., including links in source code comments [46] and even referring to research papers in repository README files [99] or source code [50]. But we still don't understand how common the practice is, how complete the mentions are, and what the supply chains of ideas look like. Nor do we have ways of systematically tracking ideas as part of the platform design, which is worth exploring.

These design principles extend naturally to AI-powered development environments, which are rapidly becoming the primary interface through which developers encounter new code and ideas. How AI coding tools are designed to surface (or suppress) informational diversity may matter as much for innovation as the social network structures we study here. We return to this point below.

Theory.

More broadly, developers have many opportunities for relatively low-effort interaction with each other and with new technologies beyond the starring behavior we study here. These include using social media to stay current [37, 83, 103], listening to podcasts [34], creating or consuming video content online [19], and participating in developer conferences [92] or trainings.

Because our study was theory-driven rather than exploratory, our findings may also be informative for thinking about these broader forms of interaction. In particular, the broader weak-ties literature suggests that low-intensity interactions can sometimes expose individuals to more diverse information environments. While we do not study these other contexts directly, it is plausible that similar patterns of informational exposure and diversity may also arise there. The frequently discussed "water-cooler effect" in workplace settings provides another example often interpreted through a similar lens: casual encounters with weaker social ties may increase opportunities for exposure to unfamiliar ideas and perspectives [12].

We need more research to investigate how different forms of low-effort interaction relate to software development outcomes, including whether some interaction types are associated with greater informational diversity or novelty than others. For example, many technology companies support their engineers regularly attending conferences and these experiences are likely beneficial. Could we quantify these benefits both in general, and relative to, say, consuming technical social media content?

Separately, our work also aligns with Baltes and Diehl's [5] theory of software development expertise. For example, the weak-tie interactions we study can be interpreted as forms of *behavior* associated with the development of a developer's knowledge base, or as instances of *continuous learning*. Similarly, GitHub can be viewed as part of the broader *work context* within which developers encounter and engage with diverse projects. But perhaps more interestingly, while the theory emphasizes *deliberate practice* as a central mechanism in expertise development, our findings may point to a complementary perspective and possible direction for theoretical refinement: casual exposure to diverse projects may also be associated with more novel forms of software recombination, beyond what focused practice alone would predict.

Exploration vs Exploitation.

Research has shown that developers are more likely to join projects that are technically more familiar to them [28, 35]. Various automated project recommendation tools have also been proposed that match projects to developers with similar technical backgrounds [107]. Such matching is likely useful, as open-source projects depend heavily on contributions from volunteers, and contributors can be in short supply [20]. At the same time, our findings suggest that maximizing technical similarity between a project and the backgrounds of its contributors may not always align with the kinds of informational diversity associated with more novel software recombinations in our data. More research is needed to better understand potential trade-offs between *exploration* and *exploitation* [61] in software development contexts, particularly in open source. For example, while computational approaches for tracking novelty are becoming increasingly feasible (as our work illustrates), we still know relatively little about when novel recombinations emerge during a project's lifecycle, under what collaborative conditions they arise, or how they relate to different forms of developer interaction. Future work could investigate how novel design patterns emerge at the onset of projects and compare these dynamics with later-stage activities such as feature-request discussions, which may provide additional opportunities for introducing new combinations of ideas and technologies over time.

Diversity. Questions around the value of having access to diverse information have also come up in research focused on demographic team composition, e.g., around gender [95] and race / ethnicity [81]. This past work posits a similar underlying mechanism (surface-level demographic variables act as a proxy for deeper-level, less observable differences in backgrounds, skills, approaches to solve problems, etc) to the one we measure here more directly. Now that such measurements are becoming possible, it would be interesting to further test the relationships between team surface-level attributes, informational or network diversity, and outcomes.

AI and Creativity.

Our dataset spans projects created before 2022, a period that largely predates the widespread adoption of LLM-based coding assistants such as GitHub Copilot (launched mid-2022) and ChatGPT (launched late 2022). This temporal boundary is important because the relationships we document between weak-tie diversity and project novelty primarily reflect developer interaction patterns in a largely pre-AI software development environment. As such, our findings should not be interpreted as evidence about how AI-assisted development affects software creativity today.

At the same time, our results raise several open questions about how AI systems may reshape informational diversity and software novelty in the future. Prior work has suggested that AI systems trained predominantly on common

patterns in existing code may encourage convergence toward frequently observed solutions, potentially reducing diversity and novelty even while improving productivity [51]. From this perspective, AI-assisted coding could weaken some of the diversity-generating mechanisms associated with weak ties.

Conversely, modern AI agents are increasingly capable of browsing repositories, synthesizing information across heterogeneous technical sources, and exposing developers to a broader range of ideas than they might otherwise encounter through their immediate social networks alone. In this sense, AI systems may also function as a new channel of informational exposure and recombination.

Determining whether AI systems ultimately increase or decrease software novelty remains an open empirical question. More broadly, our methodology may provide a useful baseline for future comparative studies examining how informational diversity and recombination patterns differ between pre-AI and post-AI software development environments. Future work could replicate our analyses on post-2022 repositories to examine whether the relationships observed here persist, weaken, or transform under AI-assisted development settings.

7 Limitations

Our study has several limitations that future work should address. First and most importantly, our findings are correlational, not causal. While we use temporal ordering to ensure that developers' network interactions precede the creation of the projects we measure, this does not rule out selection effects: developers who are inherently more creative or curious may be simultaneously more likely to star diverse repositories *and* more likely to produce innovative projects, independent of any knowledge transfer. Unobserved individual traits such as openness to experience, educational background, or prior cross-domain exposure could thus confound the observed association. Causal identification of the weak tie mechanism would ideally require an exogenous source of variation in developers' network diversity, such as a natural experiment or a platform-level intervention.

Second, our study focuses exclusively on Python projects hosted on GitHub, which may limit the generalizability of our findings. The Python ecosystem is unusually broad, spanning domains such as data science, web development, automation, scientific computing, and machine learning. This diversity may increase opportunities for developers to encounter heterogeneous technical ideas and recombine packages across otherwise distant domains, potentially amplifying the innovation benefits associated with weak ties. In contrast, ecosystems centered around narrower application domains or more rigid dependency structures may exhibit weaker effects.

At the same time, several characteristics of GitHub may also shape the dynamics we observe. GitHub strongly facilitates lightweight and low-cost interactions, such as starring repositories, following developers, and passively browsing projects, which may enhance the informational value of weak ties relative to stronger forms of collaboration. Other software ecosystems or hosting platforms that place less emphasis on public social discovery, or that rely more heavily on private organizational development, may generate different patterns of knowledge exposure and recombination.

Nevertheless, we suspect that the broader mechanism linking diverse weak-tie exposure to combinational innovation may extend beyond Python and GitHub, since similar processes have been theorized in the wider innovation and social network literature. Future work replicating our analysis across different programming languages, platforms, and organizational settings would help clarify the extent to which our findings reflect general properties of software innovation versus characteristics specific to the Python open-source ecosystem.

Third, our three interaction networks capture only a subset of the ways developers encounter and absorb new ideas. Channels such as social media, podcasts, conferences, and private communication are not reflected in our data. To the

extent that these channels also mediate knowledge diffusion, our measures of informational diversity are incomplete, and the true effect of diverse information exposure on innovation may be larger than we estimate.

Lastly, our measure of project innovativeness should also be interpreted with an important scope limitation in mind. Specifically, the atypicality metric captures innovation through novel combinations of software packages and dependencies. However, many forms of software innovation are not directly reflected in dependency recombination patterns. For example, algorithmic advances, architectural innovations, user experience (UX) improvements, and novel domain models may substantially improve a software system without necessarily introducing atypical dependency combinations. Accordingly, our measure should not be interpreted as a comprehensive representation of all forms of software innovation, but rather as a specific indicator of combinational novelty in software development. In addition, although our embedding-based atypicality metric is positively correlated with the reshuffle-based measure proposed by Fang et al. ($r = 0.65$), the correlation is moderate rather than near-perfect, suggesting that the two measures capture substantially overlapping yet partially distinct patterns of software novelty.

8 Conclusion

Much like in Granovetter's job search study exposing the strength of weak ties, we find that open-source developers may also draw useful ideas for innovative projects through passive GitHub interactions (like starring repositories) more so than through active engagement like committing to a repository. Concretely, we find that it is not the *amount* of interactions developers have with other projects and developers that predicts the innovativeness of their own projects, but rather the *diversity* of those interactions, and in particular, the extent to which that diversity is attributable to weak ties. These results hold across years, project sizes, and ownership types, and are robust to several operationalization choices.

Our study provides, to the best of our knowledge, the first large-scale empirical evidence linking the strength-of-weak-ties mechanism to innovation in open-source software. By reconstructing multi-type interaction networks at scale, operationalizing tie strength through interaction effort, and measuring project innovativeness via an embedding-based atypicality measure validated against both community star counts and expert-curated awesome lists, we offer both a methodological contribution and a new empirical instantiation of a well-established social science theory in a software engineering context.

The implications extend in several directions. For platform designers, our findings suggest that facilitating exposure to informationally diverse repositories, through diversity-aware recommendation algorithms or more prominent recognition of lurking behavior, could meaningfully support innovation in the developer community. Our dataset, which predates the widespread adoption of LLM-based coding assistants, establishes a pre-AI baseline for the human social network mechanisms we study. This raises urgent questions for future research: as AI tools increasingly mediate how developers encounter and absorb new ideas, will the innovation benefits of diverse weak ties be amplified by agents capable of accessing broad knowledge autonomously, or suppressed by assistants that steer developers toward the most common patterns? We hope this work encourages further empirical investigation into these dynamics, as the tools developers use to learn and build continue to rapidly evolve.

Data Availability. Our replication package is online DOI 10.5281/zenodo.20576356 [48]. It includes data and scripts to reproduce our tables and figures.

References

[1] Shoaib Abdul Basit and Kehinde Medase. 2019. The diversity of knowledge sources and its impact on firm-level innovation: Evidence from Germany. *European Journal of Innovation Management* 22, 4 (2019), 681–714.

[2] Abdulkareem Alali, Huzefa Kagdi, and Jonathan I Maletic. 2008. What's a typical commit? a characterization of open source software repositories. In *2008 16th IEEE international conference on program comprehension*. IEEE, 182–191.

[3] Adam Alami, Marisa Leavitt Cohn, and Andrzej Wąsowski. 2019. Why does code review work for open source software communities?. In *International Conference on Software Engineering (ICSE)*. IEEE, 1073–1083.

[4] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*. 519–528.

[5] Sebastian Baltes and Stephan Diehl. 2018. Towards a theory of software development expertise. In *International Conference on the Foundations of Software Engineering (FSE)*. 187–200.

[6] Alejandra Beghelli and Sara Jones. 2020. On the novelty of software products. In *International Conference on Design Creativity (ICDC)*. The Design Society, 11–18.

[7] Tim Berners-Lee. 1999. *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. Harper San Francisco.

[8] Kelly Blincoe, Jyoti Sheoran, Sean Goggins, Eva Petakovic, and Daniela Damian. 2016. Understanding the popular users: Following, affiliation influence and leadership on GitHub. *Information and Software Technology* 70 (2016), 30–39.

[9] Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.

[10] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489, 7415 (2012), 295–298.

[11] Hudson Borges and Marco Tulio Valente. 2018. What's in a github star? understanding repository starring practices in a social coding platform. *Journal of Systems and Software* 146 (2018), 112–129.

[12] Melanie S Brucks and Jonathan Levav. 2022. Virtual communication curbs creative idea generation. *Nature* 605, 7908 (2022), 108–112.

[13] Ronald S Burt. 2000. The network structure of social capital. *Research in organizational behavior* 22 (2000), 345–423.

[14] Ronald S Burt. 2004. Structural holes and good ideas. *Amer. J. Sociology* 110, 2 (2004), 349–399.

[15] Andrea Capiluppi, Alexander Serebrenik, and Leif Singer. 2012. Assessing technical candidates on the social web. *IEEE Software* 30, 1 (2012), 45–51.

[16] Gino Cattani and Simone Ferriani. 2008. A core/periphery perspective on individual creative performance: Social networks and cinematic achievements in the Hollywood film industry. *Organization science* 19, 6 (2008), 824–844.

[17] Antonio Cerone. 2013. Learning and activity patterns in OSS communities and their impact on software quality. *Electronic Communications of the EASST* 48 (2013).

[18] Gina N Cervetti, Carolyn A Jaynes, and Elfrieda H Hiebert. 2009. Increasing opportunities to acquire knowledge through reading. *Reading more, reading better* (2009), 79–100.

[19] Souti Chattopadhyay, Thomas Zimmermann, and Denae Ford. 2021. Reel life vs. real life: How software developers share their daily life through vlogs. In *International Conference on the Foundations of Software Engineering (FSE)*. 404–415.

[20] Jailton Coelho and Marco Tulio Valente. 2017. Why modern open source projects fail. In *International Conference on the Foundations of Software Engineering (FSE)*. 186–196.

[21] Wesley M Cohen, Daniel A Levinthal, et al. 1990. Absorptive capacity: A new perspective on learning and innovation. *Administrative science quarterly* 35, 1 (1990), 128–152.

[22] J Daniel Couger and Geoff Dengate. 1992. Measurement of creativity of IS products. In *Hawaii International Conference on System Sciences (HICSS)*, Vol. 4. IEEE, 288–298.

[23] Kevin Crowston and James Howison. 2005. The social structure of free and open source software development. (2005).

[24] Jose Ricardo da Silva, Esteban Clua, Leonardo Murta, and Anita Sarma. 2015. Niche vs. breadth: Calculating expertise over time through a fine-grained analysis. In *International Conference on Software Analysis, Evolution, and Reengineering (SANER)*. IEEE, 409–418.

[25] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social coding in GitHub: transparency and collaboration in an open software repository. In *ACM Conference on Computer Supported Cooperative Work (CSCW)*. 1277–1286.

[26] Jamal I Daoud. 2017. Multicollinearity and regression analysis. In *Journal of Physics: Conference Series*, Vol. 949. IOP Publishing, 012009.

[27] David Dekker, David Krackhardt, and Tom AB Snijders. 2019. Transitivity correlation: A descriptive measure of network transitivity. *Network Science* 7, 3 (2019), 353–375.

[28] Tapajit Dey, Andrey Karnauch, and Audris Mockus. 2021. Representation of developer expertise in open source software. In *International Conference on Software Engineering (ICSE)*. IEEE, 995–1007.

[29] Tapajit Dey, Sara Mousavi, Eduardo Ponce, Tanner Fry, Bogdan Vasilescu, Anna Filippova, and Audris Mockus. 2020. Detecting and characterizing bots that commit code. In *International Conference on Mining Software Repositories*. 209–219.

[30] Oscar Dieste, Alejandrina M Aranda, Fernando Uyaguari, Burak Turhan, Ayse Tosun, Davide Fucci, Markku Oivo, and Natalia Juristo. 2017. Empirical evaluation of the effects of experience on code quality and programmer productivity: an exploratory study. *Empirical Software Engineering* 22 (2017), 2457–2542.

[31] John Qi Dong, Weifang Wu, and Yixin Sarah Zhang. 2019. The faster the better? Innovation speed and user interest in open source software. *Information & Management* 56, 5 (2019), 669–680.

[32] Marina Du Plessis. 2007. The role of knowledge management in innovation. *Journal of knowledge management* 11, 4 (2007), 20–29.

[33] Henry Edison, Nauman Bin Ali, and Richard Torkar. 2013. Towards innovation measurement in the software industry. *Journal of systems and software* 86, 5 (2013), 1390–1407.

[34] Jeanette Engzell and Charlotte Norrman. 2023. Podcasts As A Learning Method In Engineering Education. In *51st Annual Conference of the European Society for Engineering Education*. SEFI, 398–405.

[35] Hongbo Fang, James Herbsleb, and Bogdan Vasilescu. 2023. Matching Skills, Past Collaboration, and Limited Competition: Modeling When Open-Source Projects Attract Contributors. In *International Conference on the Foundations of Software Engineering (FSE)*. 42–54.

[36] Hongbo Fang, James Herbsleb, and Bogdan Vasilescu. 2024. Novelty begets popularity, but curbs participation-a macroscopic view of the python open-source ecosystem. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–11.

[37] Hongbo Fang, Hemank Lamba, James Herbsleb, and Bogdan Vasilescu. 2022. "This is damn slick!" Estimating the impact of tweets on open source project popularity and new contributors. In *International Conference on Software Engineering (ICSE)*. 2116–2129.

[38] Chaim Fershtman and Neil Gandal. 2011. Direct and indirect knowledge spillovers: the "social network" of open-source projects. *The RAND Journal of Economics* 42, 1 (2011), 70–91.

[39] Lee Fleming, Santiago Mingo, and David Chen. 2007. Collaborative brokerage, generative creativity, and creative success. *Administrative science quarterly* 52, 3 (2007), 443–475.

[40] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science* 359, 6379 (2018), eaao0185.

[41] Marco Gerosa, Igor Wiese, Bianca Trinkenreich, Georg Link, Gregorio Robles, Christoph Treude, Igor Steinmacher, and Anita Sarma. 2021. The shifting sands of motivation: Revisiting what drives contributors in open source. In *International Conference on Software Engineering (ICSE)*. IEEE, 1046–1058.

[42] Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology* 78, 6 (1973), 1360–1380.

[43] Rajdeep Grewal, Gary L Lilien, and Girish Mallapragada. 2006. Location, location, location: How network embeddedness affects project success in open source systems. *Management science* 52, 7 (2006), 1043–1056.

[44] Wouter Groeneveld, Laurens Luyten, Joost Vennekens, and Kris Aerts. 2021. Exploring the role of creativity in software engineering. In *International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, 1–9.

[45] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[46] Hideaki Hata, Christoph Treude, Raula Gaikovina Kula, and Takashi Ishio. 2019. 9.6 million links in source code comments: Purpose, evolution, and decay. In *International Conference on Software Engineering (ICSE)*. IEEE, 1211–1221.

[47] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences* 117, 17 (2020), 9284–9291.

[48] James Evans James Herbsleb Hongbo Fang, Patrick Park and Bogdan Vasilescu. 2026. Replication package. doi:10.5281/zenodo.20576356

[49] Yan Hu, Shanshan Wang, Yizhi Ren, and Kim-Kwang Raymond Choo. 2018. User influence analysis for github developer social networks. *Expert Systems with Applications* 108 (2018), 108–118.

[50] Akira Inokuchi, Yusuf Sulistyo Nugroho, Supatsara Wattanakriengkrai, Fumiaki Konishi, Hideaki Hata, Christoph Treude, Akito Monden, and Kenichi Matsumoto. 2019. From academia to software development: publication citations in source code comments. *arXiv preprint arXiv:1910.06932* (2019).

[51] Victoria Jackson, Bogdan Vasilescu, Daniel Russo, Paul Ralph, Maliheh Izadi, Rafael Prikladnicki, Sarah D'angelo, Sarah Inman, Anielle Andrade, and André van der Hoek. 2024. The Impact of Generative AI on Creativity in Software Development: A Research Agenda. *ACM Transactions on Software Engineering and Methodology* (2024).

[52] Slinger Jansen. 2014. Measuring the health of open source software ecosystems: Beyond the scope of project health. *Information and Software Technology* 56, 11 (2014), 1508–1519.

[53] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M German, and Daniela Damian. 2016. An in-depth study of the promises and perils of mining GitHub. *Empirical Software Engineering* 21 (2016), 2035–2071.

[54] Christian Kandler, Rainer Riemann, Alois Angleitner, Frank M Spinath, Peter Borkenau, and Lars Penke. 2016. The nature of creativity: The roles of genetic factors, personality traits, cognitive abilities, and environmental sources. *Journal of Personality and Social Psychology* 111, 2 (2016), 230.

[55] Robert L Kaufman. 2013. *Heteroskedasticity in regression: Detection and correction*. Sage Publications.

[56] Brent Kitchens, Steven L Johnson, and Peter Gray. 2020. Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS Quarterly* 44, 4 (2020).

[57] Sandeep Krishnamurthy. 2006. On the intrinsic and extrinsic motivation of free/libre/open source (FLOSS) developers. *Knowledge, Technology & Policy* 18, 4 (2006), 17–39.

[58] Karim R Lakhani and Robert G Wolf. 2005. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. (2005).

[59] Hemank Lamba, Asher Trockman, Daniel Armanios, Christian Kästner, Heather Miller, and Bogdan Vasilescu. 2020. Heard it through the Gitvine: an empirical study of tool diffusion across the npm ecosystem. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 505–517.

[60] Paul Luo Li, Amy J Ko, and Andrew Begel. 2020. What distinguishes great software engineers? *Empirical Software Engineering* 25 (2020), 322–352.

[61] Lu Liu, Nima Dehmamy, Jillian Chown, C Lee Giles, and Dashun Wang. 2021. Understanding the onset of hot streaks across artistic, cultural, and scientific careers. *Nature Communications* 12, 1 (2021), 5392.

[62] Yuan Long and Keng Siau. 2007. Social network structures in open source software development teams. *Journal of Database Management (JDM)* 18, 2 (2007), 25–40.

[63] Yuan Long and Keng Siau. 2008. Impacts of social network structure on knowledge sharing in open source software development teams. *AMCIS 2008 Proceedings* (2008), 43.

[64] Yuxing Ma, Chris Bogart, Sadika Amreen, Russell Zaretzki, and Audris Mockus. 2019. World of code: an infrastructure for mining the universe of open source VCS data. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 143–154.

[65] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. 2013. Impression formation in online peer production: activity traces and personal profiles in GitHub. In *ACM Conference on Computer Supported Cooperative Work (CSCW)*. 117–128.

[66] Thorsten Merten, Bastian Mager, Paul Hübner, Thomas Quirchmayr, Barbara Paech, and Simone Bürsner. 2015. Requirements Communication in Issue Tracking Systems in Four Open-Source Projects.. In *REFSQ workshops*. 114–125.

[67] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[68] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[69] Audris Mockus, Roy T Fielding, and James D Herbsleb. 2002. Two case studies of open source software development: Apache and Mozilla. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 11, 3 (2002), 309–346.

[70] Rahul Mohanani, Prabhat Ram, Ahmed Lasisi, Paul Ralph, and Burak Turhan. 2017. Perceptions of creativity in software engineering research and practice. In *Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 210–217.

[71] Jagadeesh Nandigam, Venkat N Gudivada, and Abdelwahab Hamou-Lhadj. 2008. Learning software engineering principles using open source software. In *2008 38th Annual Frontiers in Education Conference*. IEEE, S3H–18.

[72] Mark EJ Newman. 2003. The structure and function of complex networks. *SIAM review* 45, 2 (2003), 167–256.

[73] Gang Peng, Jifeng Mu, and C Anthony Di Benedetto. 2013. Learning and open source software license choice. *Decision Sciences* 44, 4 (2013), 619–643.

[74] Huilian Sophie Qiu, Alexander Nolte, Anita Brown, Alexander Serebrenik, and Bogdan Vasilescu. 2019. Going farther together: The impact of social capital on sustained participation in open source. In *2019 ieee/acm 41st international conference on software engineering (icse)*. IEEE, 688–699.

[75] Karthik Rajkumar, Guillaume Saint-Jacques, Iavor Bojinov, Erik Brynjolfsson, and Sinan Aral. 2022. A causal test of the strength of weak ties. *Science* 377, 6612 (2022), 1304–1310.

[76] Gregorio Robles, Jesus M Gonzalez-Barahona, and Israel Herraiz. 2009. Evolution of the core team of developers in libre software projects. In *International Conference on Mining Software Repositories (MSR)*. IEEE, 167–170.

[77] AS Saabith, MMM Fareez, and T Vinothraj. 2019. Python current trend applications-an overview. *International Journal of Advance Engineering and Research Development* 6, 10 (2019).

[78] Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia* 126, 5 (2018), 1763–1768.

[79] Joseph Alois Schumpeter. 1939. Business cycles: a theoretical, historical and statistical analysis of the capitalist process. (1939).

[80] Joseph A Schumpeter and Richard Swedberg. 2021. *The theory of economic development*. Routledge.

[81] Sheik Shameer, Gema Rodríguez-Pérez, and Meiyappan Nagappan. 2023. Relationship between diversity of collaborative group members' race and ethnicity and the frequency of their collaborative contributions in GitHub. *Empirical Software Engineering* 28, 4 (2023), 83.

[82] Feng Shi and James Evans. 2023. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications* 14, 1 (2023), 1641.

[83] Leif Singer, Fernando Figueira Filho, and Margaret-Anne Storey. 2014. Software engineering at the speed of light: how developers stay current using Twitter. In *International Conference on Software Engineering (ICSE)*. 211–221.

[84] Param Vir Singh and Corey Phelps. 2013. Networks, social influence, and the choice among competing innovations: Insights from open source software licenses. *Information Systems Research* 24, 3 (2013), 539–560.

[85] Param Vir Singh, Yong Tan, and Vijay Mookerjee. 2011. Network effects: The influence of structural capital on open source project success. *Mis Quarterly* (2011), 813–829.

[86] Sabine Sonnentag. 1995. Excellent software professionals: Experience, work activities, and perception by peers. *Behaviour & Information Technology* 14, 5 (1995), 289–299.

[87] Sindre Sorhus. 2025. Collection of Awesome GitHub Repositories. https://github.com/sindresorhus/awesome.

[88] KR Srinath. 2017. Python–the fastest growing programming language. *International Research Journal of Engineering and Technology* 4, 12 (2017), 354–357.

[89] Yong Tan, Vijay Mookerjee, and Param Singh. 2007. Social capital, structural holes and team composition: Collaborative networks of the open source software community. *ICIS 2007 Proceedings* (2007), 155.
[90] Marco Tortoriello, Bill McEvily, and David Krackhardt. 2015. Being a catalyst of innovation: The role of knowledge diversity and network closure. *Organization Science* 26, 2 (2015), 423–438.
[91] Asher Trockman, Shurui Zhou, Christian Kästner, and Bogdan Vasilescu. 2018. Adding sparkle to social coding: An empirical study of repository badges in the npm ecosystem. In *International Conference on Software Engineering (ICSE)*. 511–522.
[92] Kimberly Truong, Courtney Miller, Bogdan Vasilescu, and Christian Kästner. 2022. The unsolvable problem or the unheard answer? A dataset of 24,669 open-source software conference talks. In *International Conference on Mining Software Repositories (MSR)*. 348–352.
[93] Scott F Turner, Richard A Bettis, and Richard M Burton. 2002. Exploring depth versus breadth in knowledge management strategies. *Computational & Mathematical Organization Theory* 8 (2002), 49–73.
[94] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. *Science* 342, 6157 (2013), 468–472.
[95] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. 2015. Gender and tenure diversity in GitHub teams. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3789–3798.
[96] Kristina P Vatcheva, MinJae Lee, Joseph B McCormick, and Mohammad H Rahbar. 2016. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)* 6, 2 (2016), 227.
[97] Georg Von Krogh, Sebastian Spaeth, and Karim R Lakhani. 2003. Community, joining, and specialization in open source software innovation: a case study. *Research Policy* 32, 7 (2003), 1217–1241.
[98] Binglu Wang, Zhengnan Ma, Dashun Wang, and Brian Uzzi. 2025. Peer Review and the Diffusion of Ideas. *arXiv preprint arXiv:2507.11825* (2025).
[99] Supatsara Wattanakriengkrai, Bodin Chinthanet, Hideaki Hata, Raula Gaikovina Kula, Christoph Treude, Jin Guo, and Kenichi Matsumoto. 2022. GitHub repositories with links to academic papers: Public access, traceability, and evolution. *Journal of Systems and Software* 183 (2022), 111117.
[100] Mairieli Wessel, Bruno Mendes De Souza, Igor Steinmacher, Igor S Wiese, Ivanilton Polato, Ana Paula Chaves, and Marco A Gerosa. 2018. The power of bots: Characterizing and understanding bots in OSS projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.
[101] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.
[102] Jing Wu, Khim-Yong Goh, and Qian Tang. 2007. Investigating success of open source software projects: A social network perspective. *ICIS 2007 Proceedings* (2007), 105.
[103] Marvin Wyrich and Justus Bogner. 2024. Beyond Self-Promotion: How Software Engineering Research Is Discussed on LinkedIn. In *International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. 85–95.
[104] Xuan Yang, Daning Hu, and Davison M Robert. 2013. How microblogging networks affect project success of open source software development. In *2013 46th Hawaii International Conference on System Sciences*. IEEE, 3178–3186.
[105] Xuan Yang, Xiao Li, Daning Hu, and Harry Jiannan Wang. 2021. Differential impacts of social influence on initial and sustained participation in open source software projects. *Journal of the Association for Information Science and Technology* 72, 9 (2021), 1133–1147.
[106] Robert B Zajonc. 1968. Attitudinal effects of mere exposure. *Journal of personality and social psychology* 9, 2p2 (1968), 1.
[107] Xunhui Zhang, Tao Wang, Gang Yin, Cheng Yang, Yue Yu, and Huaimin Wang. 2017. Devrec: a developer recommendation system for open source repositories. In *Mastering Scale and Complexity in Software Reuse: 16th International Conference on Software Reuse, ICSR 2017, Salvador, Brazil, May 29-31, 2017, Proceedings 16*. Springer, 3–11.

A Rationale for Restricting the Regression Analysis to Projects with Sufficient Cross-Project Interactions

In Section 4.2, we exclude projects whose core developers interacted with no more than one project during the 12 months preceding the creation of the focal project. This filtering procedure substantially reduces the sample size. Our readers may wonder why we do not instead assign these projects a diversity value of -1 (the minimum possible value in our diversity metric), thereby assuming an absence of diversity in knowledge flow.

We argue that this alternative specification substantially misrepresents the diversity of information flows, particularly in weak-tie networks such as the star network. Figure 6 compares the distribution of project diversity in the star network between (1) the sample of projects with sufficient star activity, which is used in the diversity analysis reported in Table 3, and (2) the full sample in which projects with insufficient star activity are assigned a diversity value of -1. Among projects with sufficient star activity, the observed star network often connects to a wide range of distant projects.

Therefore, it is unlikely that assigning a diversity value of -1 accurately reflects the actual diversity of projects with which the core developers of projects lacking sufficient star activity interact.

A more plausible interpretation is that these developers simply do not use the "star" feature frequently and instead rely on alternative channels for weak-tie interactions.

Therefore, while excluding such projects from the regression analysis may introduce some degree of selection bias, we consider this filtering necessary. Specifically, it ensures that the retained sample consists of projects whose core developers actively use all three interaction mechanisms (i.e., commits, issues, and stars), thereby enabling a more meaningful comparison of effect sizes across different types of interactions.

B Using transitivity to validate the 'strength' of different networks

In Section 4.3, we reason that interactions via commits, issues, and stars reflect ties of decreasing strength, and here we provide an empirical validation of this assumption through the comparison of transitivity among the three networks.

Transitivity measures the density of local connections in the local network [72]. Consider the illustration in Figure 7: if node A is connected to B and C, a triad is formed. In social network theory, if A's connections to B and C are both strong ties, the triad is unstable and will likely evolve into a triangle where A, B, and C are all connected. This evolution occurs because B and C have more opportunities to interact and increase their mutual familiarity due to their strong ties with A. Thus, networks characterized by strong ties typically exhibit a greater number of triangles, as opposed to triads, compared to those with weak ties [42]. Formally, this can be computed using the measurement of transitivity, defined as $T = 3 * N_{\text{triangles}}/N_{\text{triads}}$. A higher transitivity value (number of triangles relative to triads) indicates a network of stronger ties.

Similarly, we compare the transitivity values of our three networks, summarized in Table 4.[8] Overall, we observe approximately an order of magnitude (10×) difference in transitivity values between each pair of networks. The commit network displays the highest levels of transitivity, followed by the issue network, while the star network exhibits the lowest level.

Given that each network exhibits a different level of density, a factor that is often correlated with network transitivity [27], we further compared the observed networks with corresponding random graph models. Specifically, for each interaction network, we generated ten Erdős–Rényi random networks while preserving the original number of nodes and edges (and therefore maintaining the same network density). We then calculated the transitivity z-score using the following equation: $Z = \frac{Obs_T - Mean(T)}{Std(T)}$, where Obs_T denotes the transitivity of the empirical network, and $Mean_T$ and Std_T represent the mean and standard deviation of the transitivity values obtained from the ten random networks, respectively. This metric quantifies the extent to which the transitivity of the empirical network deviates from that expected under random conditions, measured in units of standard deviation. We observe that the relative ranking of the transitivity z-scores across the three networks is consistent with that of their original transitivity values.

Therefore, these findings are consistent with our theoretical understanding of tie strength.

C Replication with the atypicality measure by Fang et al.

We further test the robustness of our regression results by replicating the analysis presented in Table 3 using the original measurement of atypicality defined by Fang et al. [36] as the outcome variable. The results of this robustness analysis

[8] Since network transitivity is usually defined for undirected networks, we first convert our three networks (which are all directed) to their undirected form. That is, there is an undirected edge between nodes A and B if there is either an edge from A to B or from B to A. In addition, the reported number of nodes excludes isolates (i.e., nodes without edges) from each network, and the number of edges is calculated in the undirected version of each network.

Table 4. Validating the relatively decreasing strength of commit, issue, and star network ties.

Interaction	#Nodes	#Edges	Transitivity ($\times 10^{-2}$)	Z-score ($\times 10^{4}$)
Commits	763, 062	1, 926, 978	30.04	25.27
Issues	278, 945	727, 255	3.42	0.90
Stars	480, 394	3, 658, 543	0.23	0.23

are reported in Table 5. It is important to note that the range of the atypicality score in the main analysis differs from that in the robustness analysis. As a result, the magnitudes of the estimated coefficients may vary; however, the overall qualitative patterns remain consistent.

Table 5. Regression analysis on factors associated with differences in project innovativeness, with innovativeness measured with the atypicality measure by Fang et al. [36]

	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{ave} ($\mathbf{H}_1$)	-0.055^{***}	-0.009^{*}		-0.017^{***}
	(0.0015)	(0.0044)		(0.0045)
$Deg_{weakness}$	0.009^{***}	0.012^{*}		-0.020^{***}
	(0.0022)	(0.0054)		(0.0059)
Div_{ave} ($\mathbf{H}_2$)			0.042^{***}	0.048^{***}
			(0.0042)	(0.0044)
$Div_{weakness}$ ($\mathbf{H}_3$)			0.045^{***}	0.053^{***}
			(0.0056)	(0.0060)
Controls				
Org_owned	0.127^{***}	0.071^{***}	0.074^{***}	0.076^{***}
	(0.0058)	(0.0118)	(0.0116)	(0.0118)
$N_{\text{Owner_Stars}}$ (log)	0.017^{***}	0.011^{***}	0.010^{***}	0.013^{***}
	(0.0020)	(0.0028)	(0.0027)	(0.0028)
$N_{\text{Core_Devs}}$ (log)	0.193^{***}	0.093^{***}	0.087^{***}	0.094^{***}
	(0.0053)	(0.0114)	(0.0112)	(0.0114)
N_{Packages} (log)	0.219^{***}	0.150^{***}	0.153^{***}	0.154^{***}
	(0.0031)	(0.0080)	(0.0079)	(0.0080)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Observations	589,999	37,451	37,451	37,451
Adjusted R^2	0.043	0.034	0.038	0.038

*p<0.05; **p<0.01; ***p<0.001

D Validating our Findings with Direct Measurement of Network Degree and Diversity

In Section 5, we examined the effects of degree centrality and network diversity across different networks by decomposing these variables using principal component analysis (PCA). To further validate our findings, we conducted the same regression analysis using the direct measurements of degree centrality and network diversity from each network. The results are presented in Table 7.

A potential concern in this analysis is that correlations among network metrics derived from different networks may bias coefficient estimates and inflate their standard errors due to multicollinearity [96]. As shown in Table 6, the

degree centrality and diversity measures in the commit and issue networks exhibit moderate correlations (Pearson's r=0.55 for degree centrality and r=0.42 for diversity). In addition, degree centrality in the star network is moderately correlated with diversity measures (r=0.42). No other pair of variables of interest (i.e., degree centrality and diversity measures) exhibits a Pearson correlation exceeding 0.3. Furthermore, variance inflation factor (VIF) analysis indicates no substantial multicollinearity concerns, as all variables have VIF values below 2.

To mitigate the influence of correlated variables and obtain more reliable estimates of the direct effects of degree centrality and diversity, we exclude highly correlated variables in Models I–III. Specifically, Model I excludes the degree centrality and diversity variables from the commit network to isolate the effects of the issue and star networks. Model II excludes the corresponding variables from the issue network to capture the effects of the commit and star networks, and Model III excludes the variables from the issue network, as well as degree centrality in the star network. For completeness, Model IV includes all direct measurement variables, while acknowledging that some coefficient estimates in this specification may be less reliable due to multicollinearity.

Across all four models, degree centrality in most networks exhibits a negative relationship with project atypicality, providing limited support for $\mathbf{H}_1$. In contrast, diversity in the star network consistently shows a positive association with project atypicality, with a larger effect size than the diversity measures in the commit and issue networks. These findings provide additional support for $\mathbf{H}_2$ and $\mathbf{H}_3$. Moreover, Models I, II, and IV suggest that diversity in the commit network, which represents the strongest ties among the networks examined, may exert a stronger influence than diversity in the issue network. This pattern points to a potential curvilinear relationship between network tie strength and project atypicality, which we leave for future research.

Table 6. Correlation matrix between independent variables

	Deg_{commit}	Deg_{issue}	Deg_{star}	Div_{commit}	Div_{issue}	Div_{star}	Org_owned	$N_{\text{Owner_Star}}$	$N_{\text{Core_Dev}}$	N_{Packages}
Deg_{commit}	1.00	0.55	0.10	0.14	-0.04	-0.07	0.27	0.26	0.29	0.03
Deg_{issue}	0.55	1.00	0.15	0.19	0.18	0.01	0.18	0.20	0.18	0.06
Deg_{star}	0.10	0.15	1.00	0.11	0.13	0.42	-0.05	0.06	0.00	0.02
Div_{commit}	0.14	0.19	0.11	1.00	0.42	0.16	0.00	0.01	0.06	-0.04
Div_{issue}	-0.04	0.18	0.13	0.42	1.00	0.20	-0.08	0.01	-0.04	-0.02
Div_{star}	-0.07	0.01	0.42	0.16	0.20	1.00	-0.06	-0.03	-0.04	-0.03
Org_owned	0.27	0.18	-0.05	0.00	-0.08	-0.06	1.00	0.08	0.43	0.16
$N_{\text{Owner_Star}}$	0.26	0.20	0.06	0.01	0.01	-0.03	0.08	1.00	-0.04	-0.02
$N_{\text{Core_Dev}}$	0.29	0.18	0.00	0.06	-0.04	-0.04	0.43	-0.04	1.00	0.26
N_{Packages}	0.03	0.06	0.02	-0.04	-0.02	-0.03	0.16	-0.02	0.26	1.00

Note: All variables are log-scaled except for Org_owned, Div_{commit}, Div_{issue} and Div_{star}

Table 7. Regression analysis with direct measurements of network degree centrality and diversity

	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{commit} (log, scaled)		-0.002^{**}	-0.003^{***}	-0.004^{***}
		(0.0008)	(0.0008)	(0.0009)
Deg_{issue} (log, scaled)	0.002^{*}			0.003^{***}
	(0.0008)			(0.0009)
Deg_{star} (log, scaled)	-0.006^{***}	-0.006^{***}		-0.006^{***}
	(0.0008)	(0.0008)		(0.0008)
Div_{commit} (scaled)		0.006^{***}	0.005^{***}	0.006^{***}
		(0.0007)	(0.0007)	(0.0008)
Div_{issue} (scaled)	$0.001^{n.s.}$			-0.001
	(0.0007)			(0.0008)
Div_{star} (scaled)	0.011^{***}	0.010^{***}	0.007^{***}	0.010^{***}
	(0.0008)	(0.0008)	(0.0007)	(0.0008)
Controls				
Org_owned	0.017^{***}	0.018^{***}	0.019^{***}	0.018^{***}
	(0.0016)	(0.0016)	(0.0016)	(0.0016)
$N_{\text{Owner_Stars}}$ (log)	0.0008^{*}	0.001^{***}	0.001^{**}	0.001^{**}
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
$N_{\text{Core_Devs}}$ (log)	0.018^{***}	0.018^{***}	0.018^{***}	0.018^{***}
	(0.0015)	(0.0016)	(0.0016)	(0.0016)
N_{Packages} (log)	0.024^{***}	0.024^{***}	0.023^{***}	0.024^{***}
	(0.0011)	(0.0011)	(0.0011)	(0.0011)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Observations	37,451	37,451	37,451	37,451
Adjusted R^2	0.064	0.065	0.064	0.065

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant

E Replication with Project Atypicality Computed within One-Year After First Commit

In our main analysis reported in Table 3, project atypicality is computed based on all package imports observed up to the end of the observation period (i.e., the last day of 2021), whereas the independent variables are measured prior to project creation. Consequently, differences in project atypicality may partly reflect factors emerging after project creation that are not captured by our model, especially for projects that are created long before.

To further examine this concern, we recompute project atypicality using only the packages imported within one year following each project's first commit and replicate the regression analysis presented in Table 3, and Table 5. Consistent with this new specification, the numbers of core developers and imported packages are also calculated using only commits made within one year of the first commit. The results are reported in Table 8 and Table 9. Projects with one or fewer imported packages during this period are excluded because atypicality scores are undefined in such cases, resulting in a reduced sample size. Additionally, to facilitate comparison of the network effect sizes between the one-year atypicality measure and the full-period atypicality measure, we standardize the atypicality score in Table 8 and Table 9, and reproduce the result in Table 3 and Table 5 with Table 10 and Table 11 using standardized atypicality.

The results in Table 8 and Table 9 further support our main conclusions. Greater diversity in projects' network access—particularly within weak-tie network structures (e.g., star networks)—remains positively associated with project atypicality, as measured by package imports within one year of the first commit. In contrast, network degree centrality continues to exhibit either null or negative effects.

Moreover, a comparison of the effect sizes of the network measures between Table 8 and Table 10, as well as between Table 9 and Table 11, indicates that the average effect of network diversity — and particularly the diversity effect associated with weak networks — on full-period (long-term) project atypicality is greater than its effect on one-year (short-term) atypicality. These findings suggest that exposure to diverse networks may exert a cumulative influence on the creation of novel projects over time, as developers progressively absorb and integrate heterogeneous ideas throughout the project development process.

Table 8. Regression analysis with atypicality (scaled) computed on package imports within one-year after project's first commit

	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{ave} ($\mathbf{H}_1$)	0.003^{**}	$-0.007^{n.s.}$		-0.014^{***}
	(0.0011)	(0.0044)		(0.0045)
$Deg_{weakness}$	$0.0005^{n.s.}$	$-0.004^{n.s.}$		-0.033^{***}
	(0.0016)	(0.0054)		(0.0060)
Div_{ave} ($\mathbf{H}_2$)			0.037^{***}	0.046^{***}
			(0.0042)	(0.0044)
$Div_{weakness}$ ($\mathbf{H}_3$)			0.036^{***}	0.049^{***}
			(0.0056)	(0.0061)
Controls				
Org_owned	0.170^{***}	0.134^{***}	0.139^{***}	0.139^{***}
	(0.0042)	(0.0115)	(0.0113)	(0.0116)
$N_{\text{Owner_Stars}}$ (log)	0.018^{***}	0.007^{*}	0.006^{*}	0.008^{**}
	(0.0015)	(0.0028)	(0.0027)	(0.0028)
$N_{\text{Core_Devs}}$ (log, 1-Year)	0.077^{***}	0.046^{***}	0.047^{***}	0.050^{***}
	(0.0040)	(0.0123)	(0.0122)	(0.0122)
N_{Packages} (log, 1-Year)	0.269^{***}	0.128^{***}	0.116^{***}	0.130^{***}
	(0.0023)	(0.0087)	(0.0087)	(0.0087)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Observations	550,202	33,547	33,547	33,547
Adjusted R^2	0.062	0.096	0.099	0.100

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant
∗Atypicality is scaled to mean of zero and standard deviation of one

F Replication with Core Developer Fixed Effects

The attributes of a project's core developers may act as important confounding factors in our analysis. To further validate our results, we add fixed effects for each core developer who appears at least twice in the dataset and replicate the regressions for Models II, III, and IV in Table 3. This validation not only controls for other potential developer-level confounding factors but also mitigates project-level omitted variables such as differences in project types, since each developer typically works on only a subset of project categories. The results are reported in Table 12.

We do not replicate Model I because including tens of thousands of fixed-effect variables makes the estimation computationally expensive. Model II uses the same specification but is more computationally efficient.

Models I, II, and III in Table 12 correspond to Models II, III, and IV in Table 3. The results show that degree centrality has either no relationship or a negative relationship with project atypicality, while average diversity remains positively

Table 9. Regression analysis with atypicality (scaled) computed on package imports within one-year after project's first commit, using Fang et al. [36]'s measurement of atypicality

	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{ave} ($\mathbf{H}_1$)	-0.036^{***}	$-0.008^{n.s.}$		-0.014^{**}
	(0.0011)	(0.0046)		(0.0047)
$Deg_{weakness}$	0.005^{**}	0.013^{*}		-0.020^{**}
	(0.0016)	(0.0056)		(0.0062)
Div_{ave} ($\mathbf{H}_2$)			0.039^{***}	0.045^{***}
			(0.0044)	(0.0046)
$Div_{weakness}$ ($\mathbf{H}_3$)			0.051^{***}	0.058^{***}
			(0.0058)	(0.0063)
Controls				
Org_owned	0.100^{***}	0.076^{***}	0.078^{***}	0.080^{***}
	(0.0043)	(0.0120)	(0.0117)	(0.0120)
$N_{\text{Owner_Stars}}$ (log)	0.009^{***}	0.010^{***}	0.009^{**}	0.011^{***}
	(0.0015)	(0.0029)	(0.0028)	(0.0029)
$N_{\text{Core_Devs}}$ (log, 1-Year)	0.161^{***}	0.092^{***}	0.091^{***}	0.095^{***}
	(0.0040)	(0.0128)	(0.0127)	(0.0128)
N_{Packages} (log, 1-Year)	0.142^{***}	0.143^{***}	0.145^{***}	0.146^{***}
	(0.0023)	(0.0090)	(0.0090)	(0.0090)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Observations	549,068	33,547	33,547	33,547
Adjusted R^2	0.037	0.027	0.031	0.031

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant
∗Atypicality is scaled to mean of zero and standard deviation of one

associated with project atypicality. The diversity loading on weak-tie networks is only marginally significant, reaching the 90% significance level in Model II and the 95% significance level in Model III.

To further examine the role of weak ties, especially the star network, we introduce Model IV. Instead of using the PCA-based diversity decomposition in Model II, Model IV directly measures diversity in the star and commit networks as independent variables, while excluding the highly correlated diversity measure in the issue network following Section D. After standardizing the diversity measures, we find that diversity in the star network is positively associated with project atypicality, with a larger effect size than diversity in the commit network. This result further supports the strong influence of weak-tie diversity, even after controlling for core developer fixed effects.

G Replication with Alternative Definitions of Core Developers and Interaction Periods

In Section 4.3, we constructed knowledge networks between projects based on interactions among their *core developers* within a *12-month* period prior to their first contributions. To assess the robustness of our findings, we conducted replication analyses using alternative definitions of core developers and varying lengths of interaction periods, as reported in Table 3.

Following established practice [69], we alternatively identify core developers as the top contributors to a project whose cumulative commits account for at least 80% of all project commits. The corresponding regression results using this alternative definition are presented in Table 13. Similarly, we extend the interaction window from 12 to 24 months

Table 10. Regression analysis with atypicality (scaled) computed on all imported packages before year 2022

	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{ave} ($\mathbf{H}_1$)	0.006^{***}	$-0.004^{n.s.}$		-0.013^{**}
	(0.0011)	(0.0043)		(0.0044)
$Deg_{weakness}$	$-0.002^{n.s.}$	$-0.005^{n.s.}$		-0.037^{***}
	(0.0016)	(0.0052)		(0.0058)
Div_{ave} ($\mathbf{H}_2$)			0.044^{***}	0.053^{***}
			(0.0040)	(0.0043)
$Div_{weakness}$ ($\mathbf{H}_3$)			0.036^{***}	0.050^{***}
			(0.0054)	(0.0059)
Controls				
Org_owned	0.183^{***}	0.122^{***}	0.131^{***}	0.128^{***}
	(0.0041)	(0.0115)	(0.0114)	(0.0115)
$N_{\text{Owner_Stars}}$ (log)	0.016^{***}	0.005^{*}	$0.005^{n.s.}$	0.007^{*}
	(0.0014)	(0.0027)	(0.0026)	(0.0027)
$N_{\text{Core_Devs}}$ (log)	0.092^{***}	0.127^{***}	0.124^{***}	0.128^{***}
	(0.0037)	(0.0111)	(0.0109)	(0.0111)
N_{Packages} (log)	0.284^{***}	0.164^{***}	0.167^{***}	0.169^{***}
	(0.0021)	(0.0078)	(0.0078)	(0.0078)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Observations	589,999	37,451	37,451	37,451
Adjusted R^2	0.054	0.059	0.063	0.064

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant
∗Atypicality is scaled to mean of zero and standard deviation of one

prior to a core developer's first contribution to the focal project when constructing the networks. The regression results based on this extended interaction period are presented in Table 14.

We also run the same model (as model IV in Table 3) with several other combinations of core developer identification approach and length of interaction periods, and the estimated coefficient of Variables of Interest are presented in Table 15

H Relationship between awesome and atypical projects.

In Section 4.6, we present that the "awesome" projects on average are more atypical per our measurement than other projects in our sample, and we further validate this result with a regression analysis in Table 16, where we show that this relationship is not driven by confounding variables such as the project size and team size.

I Real-world examples of projects with varying level of tie diversity

In Section 4.5, we present graphs depicting the networks of projects characterized by varying levels of average diversity and relative weak tie diversity. To concretize those abstract diagrams, we also provide real-world examples in Figure 8.

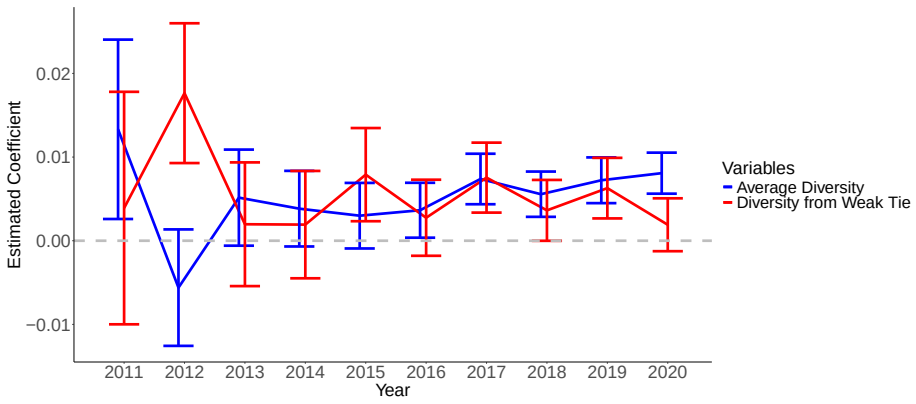
Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Table 11. Regression analysis with atypicality (scaled) computed on package imports before 2022, using Fang et al. [36]'s measurement of atypicality

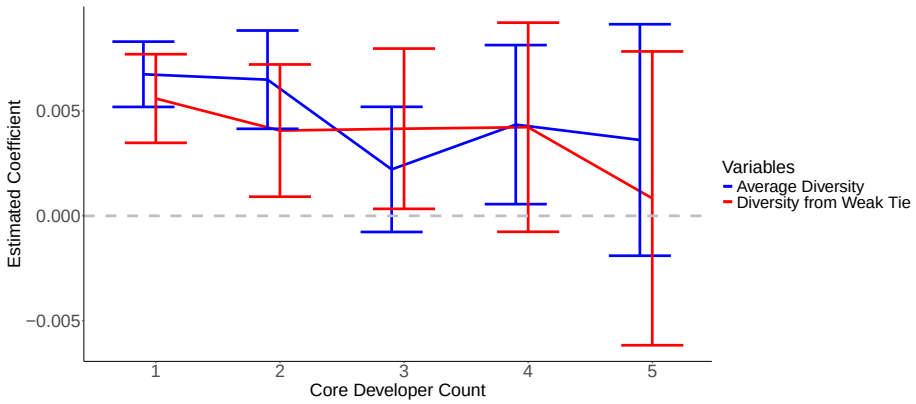
	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{ave} ($\mathbf{H}_1$)	-0.039^{***}	-0.009^{*}		-0.017^{***}
	(0.0011)	(0.0044)		(0.0044)
$Deg_{weakness}$	0.006^{***}	0.012^{*}		-0.020^{***}
	(0.0016)	(0.0053)		(0.0059)
Div_{ave} ($\mathbf{H}_2$)			0.041^{***}	0.048^{***}
			(0.0044)	(0.0043)
$Div_{weakness}$ ($\mathbf{H}_3$)			0.045^{***}	0.052^{***}
			(0.0058)	(0.0059)
Controls				
Org_owned	0.089^{***}	0.070^{***}	0.074^{***}	0.075^{***}
	(0.0041)	(0.0117)	(0.0115)	(0.0117)
$N_{\text{Owner_Stars}}$ (log)	0.012^{***}	0.011^{***}	0.010^{***}	0.013^{***}
	(0.0014)	(0.0027)	(0.0026)	(0.0027)
$N_{\text{Core_Devs}}$ (log)	0.136^{***}	0.092^{***}	0.086^{***}	0.093^{***}
	(0.0037)	(0.0113)	(0.0111)	(0.0113)
N_{Packages} (log)	0.154^{***}	0.148^{***}	0.152^{***}	0.153^{***}
	(0.0022)	(0.0079)	(0.0079)	(0.0079)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Observations	589,999	37,451	37,451	37,451
Adjusted R^2	0.043	0.034	0.038	0.038

*p<0.05; **p<0.01; ***p<0.001

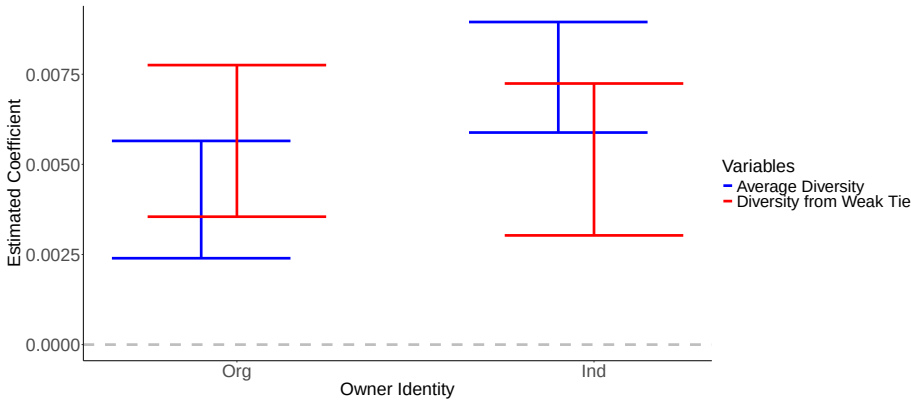
∗Atypicality is scaled to mean of zero and standard deviation of one



(a) Effects of knowledge diversity across years.



(b) Effects of knowledge diversity on different team-size projects.



(c) Effects of knowledge diversity on individual vs org projects.

Fig. 5. The effects of the two diversity of knowledge variables (Div_{ave} testing $\mathbf{H}_2$ and $Div_{weakness}$ testing $\mathbf{H}_3$) are generally robust. Error bars denote 95% confidence intervals.

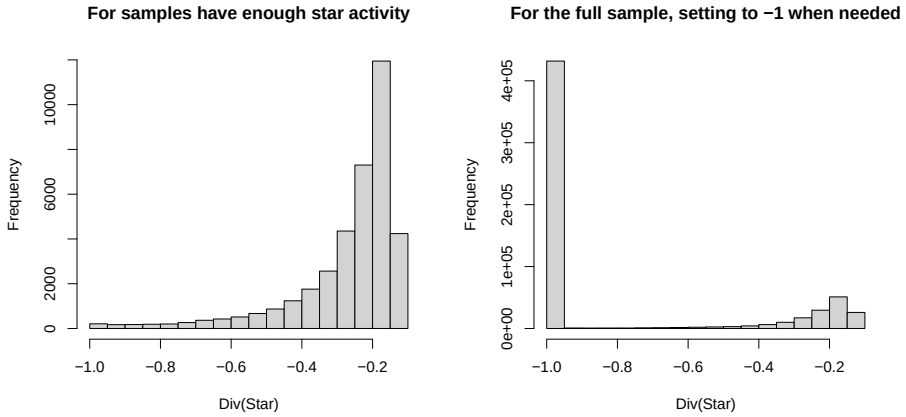


Fig. 6. The distribution of star diversity for the sample of projects whose core developers have enough star activity (left), and the full sample which set the diversity to -1 if their core developers have not starred enough projects.

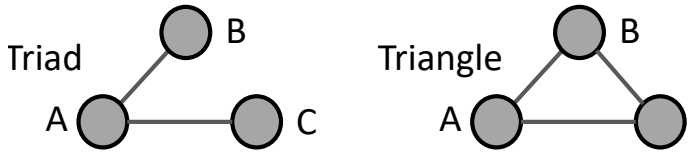


Fig. 7. In strongly tied social networks, triads are unlikely.

Table 12. Regression analysis with core developer fixed effects

	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{ave} ($\mathbf{H}_1$)	$0.001^{n.s.}$		$-0.00004^{n.s.}$	
	(0.0012)		(0.0011)	
$Deg_{weakness}$	$-0.002^{n.s.}$		-0.005^{**}	
	(0.0013)		(0.0013)	
Div_{ave} ($\mathbf{H}_2$)		0.004^{**}	0.005^{***}	
		(0.0010)	(0.0010)	
$Div_{weakness}$ ($\mathbf{H}_3$)		$0.002^{n.s.}$	0.003^{*}	
		(0.0011)	(0.0011)	
Div_{commit} (scaled)				0.003^{*}
				(0.0011)
Div_{star} (scaled)				0.004^{**}
				(0.0011)
Controls				
Org_owned	0.011^{**}	0.012^{**}	0.012^{**}	0.012^{**}
	(0.0028)	(0.0029)	(0.0027)	(0.0029)
$N_{\text{Owner_Stars}}$ (log)	0.001^{*}	0.002^{*}	0.002^{*}	0.002^{*}
	(0.0005)	(0.0005)	(0.0004)	(0.0005)
$N_{\text{Core_Devs}}$ (log)	0.018^{**}	0.018^{***}	0.018^{**}	0.018^{***}
	(0.0044)	(0.0039)	(0.0043)	(0.0040)
N_{Packages} (log)	0.029^{***}	0.030^{***}	0.030^{***}	0.030^{***}
	(0.0013)	(0.0012)	(0.0013)	(0.0013)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Core Dev	✓	✓	✓	✓
Observations	37,451	37,451	37,451	37,451
Adjusted R^2	0.195	0.196	0.196	0.196

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant

Table 13. Regression analysis with alternative core developer identification approach

	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{ave} ($\mathbf{H}_1$)	0.001^{***}	$-0.001^{n.s.}$		-0.002^{***}
	(0.0002)	(0.0006)		(0.0006)
$Deg_{weakness}$	-0.001^{**}	-0.002^{*}		-0.006^{***}
	(0.0002)	(0.0007)		(0.0008)
Div_{ave} ($\mathbf{H}_2$)			0.007^{***}	0.008^{***}
			(0.0006)	(0.0006)
$Div_{weakness}$ ($\mathbf{H}_3$)			0.004^{***}	0.006^{***}
			(0.0007)	(0.0008)
Controls				
Org_owned	0.028^{***}	0.018^{***}	0.020^{***}	0.020^{***}
	(0.0006)	(0.0016)	(0.0015)	(0.0016)
$N_{\text{Owner_Stars}}$ (log)	0.002^{***}	$0.001^{n.s.}$	$0.0004^{n.s.}$	$0.0007^{n.s.}$
	(0.0002)	(0.0004)	(0.0004)	(0.0004)
$N_{\text{Core_Devs}}$ (log)	0.010^{***}	0.017^{***}	0.015^{***}	0.016^{***}
	(0.0005)	(0.0015)	(0.0015)	(0.0015)
N_{Packages} (log)	0.043^{***}	0.025^{***}	0.025^{***}	0.026^{*}
	(0.0003)	(0.0010)	(0.0010)	(0.0010)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Observations	679,930	40,589	40,589	40,589
Adjusted R^2	0.053	0.064	0.068	0.070

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant

Table 14. Regression analysis with 24-month period length to construct networks

	Model I	Model II	Model III	Model IV
Variables of interest				
Deg_{ave} ($\mathbf{H}_1$)	0.002^{***}	$0.0003^{n.s.}$		-0.001^{**}
	(0.0002)	(0.0005)		(0.0005)
$Deg_{weakness}$	$-0.0002^{n.s.}$	-0.002^{**}		-0.007^{***}
	(0.0002)	(0.0006)		(0.0007)
Div_{ave} ($\mathbf{H}_2$)			0.007^{***}	0.008^{***}
			(0.0004)	(0.0005)
$Div_{weakness}$ ($\mathbf{H}_3$)			0.004^{***}	0.007^{***}
			(0.0007)	(0.0007)
Controls				
Org_owned	0.027^{***}	0.020^{***}	0.022^{***}	0.021^{***}
	(0.0006)	(0.0014)	(0.0014)	(0.0014)
$N_{\text{Owner_Stars}}$ (log)	0.002^{***}	0.001^{*}	0.0010^{**}	0.0012^{***}
	(0.0002)	(0.0003)	(0.0003)	(0.0003)
$N_{\text{Core_Devs}}$ (log)	0.013^{***}	0.015^{***}	0.015^{***}	0.015^{***}
	(0.0006)	(0.0014)	(0.0013)	(0.0013)
N_{Packages} (log)	0.042^{***}	0.024^{***}	0.025^{***}	0.025^{***}
	(0.0003)	(0.0009)	(0.0009)	(0.0009)
Fixed effect				
$\text{Year}_{\text{creation}}$	✓	✓	✓	✓
Observations	587,710	51,571	51,571	51,571
Adjusted R^2	0.055	0.055	0.059	0.061

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant

Table 15. Regression analysis with different core developer identification and length of interaction periods

Core developer identification	**Interaction period length (month)**	**Variables**			
		Deg_{ave}	$Deg_{weakness}$	Div_{ave}	$Div_{weakness}$
Original	6	$-0.001^{n.s.}$	-0.005^{***}	0.006^{***}	0.006^{***}
		(0.0009)	(0.0011)	(0.0008)	(0.0008)
Alternative	6	-0.003^{**}	-0.006^{***}	0.007^{***}	0.007^{***}
		(0.0008)	(0.0010)	(0.0008)	(0.0010)
Alternative	24	-0.003^{**}	-0.007^{***}	0.010^{***}	0.006^{***}
		(0.0005)	(0.0007)	(0.0005)	(0.0007)

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant

Table 16. Regression analysis on the relationship between awesome projects and project innovativeness. (Outcome variable: Project Atypicality)

	Estimated coefficient
Variables of interest	
Is awesome	0.028^{***}
	(0.0052)
Controls	
Org_owned	0.017^{***}
	(0.0016)
$N_{\text{Owner_Stars}}$ (log)	$0.001^{n.s.}$
	(0.0004)
$N_{\text{Core_Devs}}$ (log)	0.017^{***}
	(0.0015)
N_{Packages} (log)	0.023^{***}
	(0.0011)
Fixed effect	
$\text{Year}_{\text{creation}}$	✓
Observations	37,451
Adjusted R^2	0.059

*p<0.05; **p<0.01; ***p<0.001; n.s. for non-significant

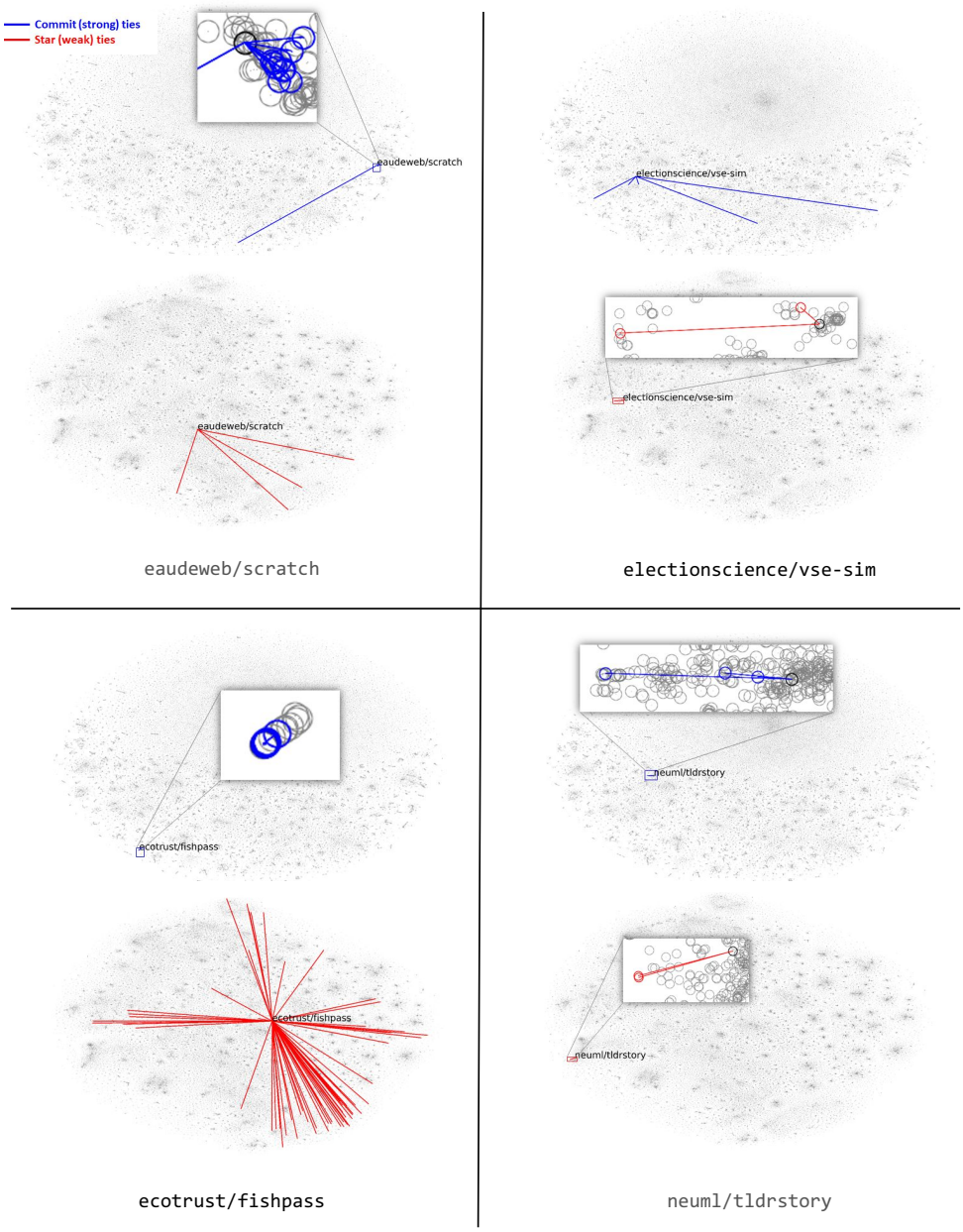


Fig. 8. Real-world examples for projects with varying level of diversity, corresponding to the four quadrants in Figure 3.