# Connected to Stay: Gender Homophily and Its Role in Open-Source Software Developer Retention

Tielin Katy Yu
Carnegie Mellon University
Pittsburgh, USA
tieliny@andrew.cmu.edu

Huilian Sophie Qiu*
Northwestern University
Evanston, USA
sophie.qiu@kellogg.northwestern.edu

Patrick Park
Carnegie Mellon University
Pittsburgh, USA
patrickp@andrew.cmu.edu

Laura Dabbish
Carnegie Mellon University
Pittsburgh, USA
dabbish@andrew.cmu.edu

Bogdan Vasilescu
Carnegie Mellon University
Pittsburgh, USA
bogdanv@andrew.cmu.edu

## Abstract

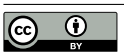
Understanding how social connections shape participation is critical for sustaining open-source software (OSS) communities. To investigate this, we analyze longitudinal collaboration networks of 1.6 million developers over 14 years (2008–2022) using the World of Code dataset, focusing on gender homophily, developers' preference to collaborate with others of the same gender, and its relationship to retention. We find strong evidence of gender homophily beyond demographic baselines, with same-gender ties among women exceeding random expectations by more than twice. While women tend to form homophilous ties, men show a preference for cross-gender collaboration, and these patterns shift over time as the OSS ecosystem evolves. Most notably, we find that developers, especially women, who collaborate with other women experience significantly lower dropout risk, with hazard rates up to 20.7% lower. These retention effects only become visible after accounting for activity levels, consistent with prior work showing that aggregate analyses can mask true relationships. Our results suggest that homophilous ties offer social support that improves developer persistence, highlighting the importance of inclusive, gender-diverse networks for sustaining participation in OSS projects.

## Keywords

Open source software, Gender homophily, Developer retention, ERGM, Survival analysis, Social network analysis, Exponential random graph models, Gender diversity

**ACM Reference Format:**
Tielin Katy Yu, Huilian Sophie Qiu, Patrick Park, Laura Dabbish, and Bogdan Vasilescu. 2026. Connected to Stay: Gender Homophily and Its Role in Open-Source Software Developer Retention. In *2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE '26), April 12–18, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3744916.3773181

*Corresponding author.




*ICSE '26, Rio de Janeiro, Brazil*

ACM ISBN 979-8-4007-2025-3/26/04
https://doi.org/10.1145/3744916.3773181

## 1 Introduction

The economic and societal value of open-source software (OSS) is now widely acknowledged [14]. OSS libraries and packages are used across sectors by corporations, nonprofits, government agencies, scientists, students, and hobbyists alike [52]. Sustained participation is critical for the maintenance and success of OSS projects [7, 15]. However, prior work has consistently shown that this effort is fragile: contributors can disengage at any time for a variety of reasons [2].

Among the many sustainability challenges OSS faces, low gender diversity stands out as especially concerning—not only as a matter of equity, but because it undermines potential benefits to team performance. A growing body of research shows that gender-diverse teams are more productive and exhibit fewer negative collaboration patterns, or "community smells" (negative patterns in team communication and collaboration) [6, 59]. Diverse teams may also be better equipped to understand and serve the diverse user base of OSS [51].

While women's underrepresentation in OSS is well documented [56, 68, 73], less is known about the social mechanisms that shape their participation and retention. Unlike formal organizations, OSS communities lack institutional retention structures and instead depend on intrinsic motivation and social ties to sustain contributions [49]. Understanding how social networks and social support affect who stays or leaves remains a critical gap in the literature.

In social sciences, homophily theory has been widely used to study social tie formation. Homophily, the idea that "similarity breeds connection" [47], is a fundamental principle of social network analysis. Previous research in many domains has shown that people tend to form social bonds with others who share similar characteristics. Among the attributes most commonly studied are race, ethnicity, age, religion, education, occupation, and sex [47, 75]. However, much of the empirical work on gender homophily has focused on formal organizational settings [33, 38]. The dynamics and implications of homophily in voluntary, project-based technical communities like OSS remain underexplored. OSS networks provide a rich empirical setting to study these questions: they combine task-driven collaboration and voluntary association across time and geography, while leaving detailed digital records of developer interactions [11].

In this study, we investigate how gender influences collaborative tie formation in OSS communities, and whether these ties, in turn,

shape developer retention. We guide our investigation with two research questions:

**RQ$_1$**. *To what extent do OSS collaborative networks exhibit gender homophily beyond baseline expectations derived from gender composition and structural constraints?*

This question examines whether observed collaboration patterns reflect social preferences rather than network topology or demographic imbalances. We assess this by comparing observed tie frequencies to null models that simulate network formation without gender-based preferences, allowing us to isolate the effect of homophily on network structure [39, 47].

**RQ$_2$**. *What patterns characterize gender-based tie formation and how do these patterns correlate with retention outcomes across developer characteristics and temporal contexts?*

Here, we move from describing patterns to uncovering mechanisms. We explore whether gender homophily operates through selective tie formation, tie persistence, or both, and how these patterns influence the likelihood of continued contribution over time [58, 65].

To answer these questions, we employ a multi-method approach grounded in longitudinal social network analysis. We analyze 14 years (2008–2022) of longitudinal collaboration data covering 1.6 million developers from the World of Code (WoC) dataset.

First, we use observed-to-expected ratio analysis to quantify gender homophily at both global and local network levels. Second, we apply egocentric exponential random graph models (ERGMs) [32] to identify the micro-level processes driving tie formation while accounting for individual and structural factors. Finally, we use Cox proportional hazards survival analysis [9] to assess how specific network configurations predict developer departure risk over time. Together, these methods allow us to unpack not just whether gender homophily exists in OSS collaboration, but also how it matters for sustaining participation.

Our analysis reveals that women form disproportionately high levels of same-gender ties, over 200% more than expected by chance, despite their numerical minority. Critically, these gender-homophilous connections are strongly associated with lower dropout risk: women with same-gender collaborators are up to 20.7% less likely to disengage, suggesting that these ties offer crucial social support in male-dominated environments. To support reproducibility, we provide a complete replication package including all analysis scripts, processed data, and documentation.[1] Our findings offer new insights into the social fabric of open-source software development. By showing how gendered collaboration patterns influence retention, this work highlights the importance of inclusive network structures for sustaining voluntary technical communities.

## 2 Related Work

### 2.1 Gender Homophily Theory

The theory of homophily, the tendency for individuals to associate with others who are similar to themselves, is a foundational concept in social network analysis, often summarized by the adage "birds of a feather flock together" [31, 47]. Social similarity based on attributes such as gender, age, race, and education has been shown to facilitate communication and relationship formation [29, 31, 60]. Homophily plays a critical role in shaping both personal relationships and professional collaborations across a range of domains.

Monge and Contractor [50] identify two main theoretical mechanisms underlying homophily: the similarity-attraction hypothesis and self-categorization theory. The similarity-attraction hypothesis suggests that individuals are more likely to interact with others who share similar observable or perceived traits [3]. Self-categorization theory, in contrast, emphasizes the cognitive processes by which individuals classify themselves and others into social categories based on attributes such as gender, race, and socio-economic status, resulting in preferential attachment to those seen as similar [1, 70].

Gender homophily, in particular, has been shown to emerge early in life. As McPherson *et al.* [47] observed, children begin to recognize gender as a stable personal characteristic before entering school, and this awareness is reflected in their social behavior. Developmental studies find that girls tend to form smaller, more intimate groups, while boys often participate in larger, more loosely structured play networks [44, 46]. These early tendencies lay the groundwork for gendered patterns of interaction and exclusion that persist into adulthood, contributing to gender segregation in workplaces and adult social networks [45, 48].

### 2.2 Homophily Formation: Preference versus Structure

A fundamental challenge in homophily research involves distinguishing genuine social preferences from structural opportunities. Rivera *et al.* [58] caution that researchers often "significantly overestimate actors' preferences to interact with self-similar others by neglecting to account for people's self-selection into contexts that bring together disproportionately homophilous actors." For example, college students with similar characteristics "tend to take classes together, where they find occasion to create social ties," suggesting that observed homophily may reflect shared contexts rather than active preferences.

In OSS contexts, this attribution problem is particularly complex. Developers may collaborate based on technical expertise and past success [27], shared project interests, or functional interdependencies [35], creating structural opportunities for gender-based patterns that may appear preference-driven. McPherson *et al.* [47] emphasize that understanding "how the organizational structure relates to the personal networks of individuals" is crucial for interpreting homophily patterns.

While women's underrepresentation in OSS is well documented [56, 68, 73], prior work has primarily examined individual-level factors (*e.g.*, bias in code review [67, 68], unwelcoming community environment [17]) or aggregate diversity metrics (*e.g.*, percentage of women contributors [56, 73]), yet little is known about why and how gender homophily forms in OSS contexts or how it relates to developer retention. Network-based approaches that examine the *relational structures*, *e.g.*, patterns of who collaborates with whom, remain rare in software engineering research, despite their proven value in organizational and health research contexts [28, 42]. This gap is particularly notable given that OSS participation depends fundamentally on voluntary collaboration and social ties rather than formal organizational structures [49]. Our work addresses

[1] https://zenodo.org/records/16119508

this gap by applying social network analysis to examine how gendered collaboration patterns shape retention outcomes, bridging homophily theory, ERGM methodology, and software engineering practice.

## 2.3 ERGM Applications Across Disciplines

ERGMs have emerged as a leading statistical framework for analyzing network formation processes across diverse fields [19, 42, 76]. Their key strength lies in the ability to model homophily and other network features while accounting for structural dependencies, making them particularly appropriate for studying social tie formation based on attribute similarity [22, 40]. ERGMs explicitly model homophily through terms like `nodematch`, allowing researchers to quantify the tendency for individuals with similar attributes to form ties beyond what would be expected by chance [32].

In organizational settings, ERGMs have been used to study intra-organizational networks including information sharing, advice seeking, and collaboration [72]. Public health researchers have adopted ERGMs to study collaboration networks among healthcare organizations [28]. ERGMs have also been applied to study knowledge diffusion in educational contexts [78]. ERGMs explicitly model homophily through terms like `nodematch`, allowing researchers to quantify the tendency for individuals with similar attributes to form ties beyond what would be expected by chance [32]. We provide a detailed introduction to ERGMs and their application to our research questions in Section 3.5.

Despite their widespread adoption across social sciences, ERGMs remain underutilized in software engineering research, representing a methodological opportunity for understanding collaboration patterns in technical communities.

# 3 Methods

Figure 1 provides an overview of our analytical approach. Our methodology follows a three-track design to comprehensively examine gender homophily in OSS collaboration networks: descriptive analysis to establish empirical patterns, inferential ERGM analysis to distinguish preference-driven from structural effects, and survival analysis to examine retention outcomes.

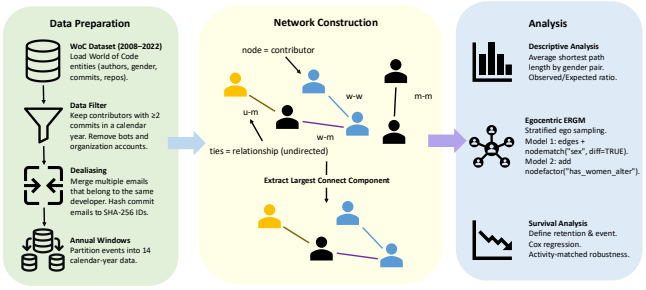



**Figure 1: Overview of our analytical pipeline**

## 3.1 Data: The World of Code Dataset

We use the World of Code (WoC) dataset, a large-scale infrastructure that mines version control data across Free/Libre Open Source Software (FLOSS) ecosystems [43]. WoC enables comprehensive cross-referencing of authors, projects, commits, and dependencies over time, supporting large-scale network analyses.

Gender classification is inferred probabilistically using origin-aware name-to-gender inference. Specifically, we use gender labels from WoC curated data, which includes NamSor V2 outputs (`likelyGender` and `probabilityCalibrated`) for name-to-gender inference [5, 43]. Following recent work [56], we adopt a 0.7 probability threshold: a contributor is classified as woman if the female probability is $\geq 0.7$, as man if the male probability is $\geq 0.7$, and as unknown otherwise. This origin-aware approach accounts for cultural variation in name-gender associations. While name-based gender inference has limitations due to pseudonymity and the 30.93% unknown classification rate, it remains the most scalable method for ecosystem-wide analysis. Large-scale name-based inference is standard practice in OSS gender research, with comparative studies placing NamSor among top performers while noting locale-specific caveats [13, 61–63]. Our binary labels reflect data constraints, not contributors' self-identification.

*3.1.1 Dataset Characteristics and Scope.* Our analysis encompasses a substantial network representing distributed software development collaboration with the following characteristics:

- **Package Ecosystems:** 38 distinct ecosystems
- **Software Projects:** 1,647,835 repositories
- **Individual Contributors:** 2,682,608 unique developers
- **Temporal Coverage:** 175 months (January 2008–June 2022)

We partition the temporal span into 14 annual observation windows, creating distinct social networks for each calendar year. This temporal granularity balances the need to capture sustained collaborative relationships while maintaining sensitivity to network evolution dynamics. Annual windows provide sufficient duration to observe meaningful collaboration patterns while enabling detection of temporal changes in network structure, following established practices in longitudinal OSS research [49].

*3.1.2 Treatment of Unknown Classifications.* Gender distribution within the network, based on Namsor classification, reveals demographic composition characteristic of OSS communities: Men (66.38%), Women (2.70%), and Unknown (30.93%). The substantial proportion of unknown classifications reflects both technical limitations of name-based inference and prevalent pseudonymous participation patterns in developer communities.

About 30.93% of developers in our dataset have unknown gender classifications. Prior work suggests these users exhibit behavioral patterns closer to women than men [74]. To address this, we include unknowns in shortest path distance calculations—potentially dampening homophily effects, while restricting primary homophily analyses to identified men and women for consistency, following established OSS gender research practices [68]. The proportion of identified women (2.70%) aligns with previous OSS demographic studies [56], supporting the representativeness of our data.

## 3.2 Network Construction and Definition

For each calendar year, we constructed a social network to map the relationships among contributors within the OSS community. In this network, each *node* represents an individual contributor. A

*tie* is a collaborative relationship between two contributors. A tie is established between two users if they have collaborated on the same project within the same annual observation window. The network is *undirected*, reflecting mutual collaboration. A tie's weight corresponds to the total number of projects the two collaborators have jointly worked on, *i.e.*, if 2 people have jointly worked on 2 open-source projects in a given period, there is a tie between them with a weight equals 2.

*3.2.1 Defining Collaboration.* In this study, collaboration means working together to achieve a common goal within a software project. This definition does not require direct interaction between contributors, but rather represents the *opportunity structure* for direct interaction and shared focus. A tie is established when two contributors make commits to the same repository within the same annual window, indicating co-participation in a common project context. We acknowledge that this operationalization is broader than definitions based on specific micro-activities (*e.g.*, co-commits on the same file, co-authoring pull requests, or exchanging comments on issues [23]). However, this approach is appropriate for our research context for three reasons:

First, the dominant mode of software production on GitHub cannot be narrowly characterized by joint participation in micro-activities alone. Even when two developers do not directly interact, they share a common focus of attention on project activities that may not directly involve them, but still shape their perceptions, attitudes, and subsequent interactions – factors that likely influence retention.

Second, project participation serves as an important signal and source of social identity for developers, as evidenced by the prominent display of projects on GitHub profiles [12].

Third, developer interactions frequently occur outside GitHub on platforms like Twitter/X or Slack [64]. We lack data on these external interactions, but we can reasonably assume they are more likely when developers collaborate on multiple projects together. While external interactions are beyond our study's scope, our operationalization of collaboration likely approximates these unobserved interactions better than narrower definitions would.

Additionally, we filtered out dyads with only one shared project in a year to ensure that incidental co-occurrence does not create an edge, requiring sustained collaboration across multiple projects for a tie to exist. This multi-project requirement strengthens our confidence that observed ties represent meaningful collaborative relationships rather than random co-presence.

*3.2.2 Network Filtering.* We applied filtering to focus on sustained collaborative relationships. Contributors who made only a single commit across all projects during a calendar year were excluded from the analysis. This threshold removes transient or one-time contributors who are unlikely to have established meaningful collaborative relationships, allowing us to focus on developers with at least minimal sustained engagement in the ecosystem. This filtering is consistent with prior work studying active OSS contributors [8, 36].

*3.2.3 User Identification.* Contributors were uniquely identified using SHA-256 encoded email addresses from commit metadata. We performed additional dealiasing procedures to consolidate multiple email addresses belonging to single developers, ensuring accurate contributor identification and preventing artificial network fragmentation from users with multiple email accounts [43]. Each consolidated developer identity corresponds to one node in the network.

*3.2.4 Connected Component Extraction.* Following contributor filtering, we extract the largest connected component (LCC) from each annual network, defined as the subnetwork with the most nodes (contributors) that are connected to each other through some path. This follows standard network science practice [16, 53, 55], as the LCC represents the cohesive core where homophily can meaningfully influence mixing and persistence. Nodes in the LCC have shorter path distances, greater potential reach and influence, and collectively reveal the network's overall connectivity patterns. By focusing on the LCC, we analyze the portion of the network where gender-based collaboration patterns have sufficient structural context for reliable modeling, while avoiding isolates and very small components that lack the structural variation needed for stable parameter estimates [32, 55].

## 3.3 Descriptive Analysis: Network Structure and Gender Patterns

To provide initial evidence addressing $RQ_1$ and establish the empirical foundation for our subsequent modeling approaches, we conduct a descriptive analysis of OSS collaboration networks across all 14 observation years. This analysis characterizes fundamental network properties, temporal evolution patterns, and gender-based structural differences prior to implementing targeted sampling for ERGM and survival analyses.

*3.3.1 Network Scale and Temporal Evolution.* We examine the structural evolution of OSS collaboration networks across our 14-year observation period (2008-2022) to understand the ecosystem's growth patterns and demographic composition stability.

*Network Growth Dynamics*: We tracked temporal changes in network scale, measuring the number of active contributors (nodes) and collaborative relationships (edges) within each calendar year.

*Gender Representation Analysis*: We analyzed the gender composition of each annual network, calculating the proportion of contributors classified as women, men, and unknown gender. This temporal analysis established whether gender representation patterns remained stable or evolved systematically during the observation period.

*3.3.2 Shortest Path Distance Analysis.* To characterize the structure of collaborative relationships and examine gender-based patterns in network connectivity, we analyze shortest path distances between different types of gender pairings within each year. This analysis provides insights into how closely connected contributors of different gender combinations tend to be within the OSS collaboration structure.

The shortest path distance represents the minimum number of collaborative steps needed to connect two developers through the network. We calculate these distances for woman-woman, man-man, and woman-man dyads using breadth-first search algorithms, which systematically explore the network level by level to find the

shortest routes between all developer pairs. All network computations were performed using the EasyGraph library [20], which provides efficient implementations for large-scale network analysis.

*3.3.3 Observed-to-Expected Ratio Analysis.* To distinguish between preference-driven homophily and gender-induced similarity patterns, we implement observed-to-expected (O/E) ratio analysis that controls for gender representation imbalances while testing whether collaboration patterns deviate from random mixing expectations.

*Expected Tie Calculation*: For each observation window $t$, we calculate expected tie frequencies under a null model of random mixing that preserves total edge count while randomizing gender-based tie distribution. Let $N_w^{(t)}$ and $N_m^{(t)}$ denote the number of women and men respectively in window $t$, with total population $N^{(t)} = N_w^{(t)} + N_m^{(t)}$ and $E^{(t)}$ observed edges. Under random mixing, the baseline connection probability is $P_{connect}^{(t)} = \frac{2E^{(t)}}{N^{(t)}(N^{(t)}-1)}$, representing network density. Expected ties are:

$$E[T_{ww}^{(t)}] = \binom{N_w^{(t)}}{2} P_{connect}^{(t)}$$

$$E[T_{mm}^{(t)}] = \binom{N_m^{(t)}}{2} P_{connect}^{(t)}$$

$$E[T_{wm}^{(t)}] = N_w^{(t)} N_m^{(t)} P_{connect}^{(t)}$$

where binomial coefficients count possible same-gender pairs.

*Multi-Distance Analysis*: We calculate O/E ratios across three network distances: path length 1 (direct collaborations testing immediate partnership preferences), path length 2 (two-step connections examining local neighborhood clustering), and path length 3 (three-step connections assessing broader structural positioning).

*Ratio Computation*: For each gender pairing $g$, distance $d$, and window $t$, we compute

$$O/E_{g,d}^{(t)} = \frac{T_{observed}(g,d,t)}{E[T_{g,d}^{(t)}]}$$

where $T_{observed}(g,d,t)$ represents observed connections at distance $d$ and $E[T_{g,d}^{(t)}]$ denotes expected counts under random mixing. Ratios above 1.0 indicate positive homophily; ratios below 1.0 suggest structural constraints or heterophily [47, 53].

## 3.4 Stratified Sampling to Obtain Egocentric Networks for ERGM and Survival Analysis

Given the computational complexity of analyzing complete networks containing millions of nodes and the need to ensure adequate representation of rare collaboration patterns, we implement a stratified egocentric sampling approach [41, 55]. *Sampling Rationale*: The complete OSS networks are too large for direct analysis using ERGMs or detailed survival models. Additionally, the low proportion of woman contributors (2.70%) means that woman-woman collaborative relationships would be severely underrepresented in random sampling approaches.

*Stratification Framework*: For each annual observation window, we implement stratified sampling across four distinct categories that capture gender-collaboration patterns:

(1) Women contributors who have at least one woman collaborator in their network
(2) Women contributors whose collaborative networks contain no other women
(3) Men contributors who have at least one woman collaborator
(4) Men contributors whose networks contain no woman collaborators

*Sample Size Determination*: We determine the base sample size using the smallest stratum (typically women with ≥1 women collaborator), then sample equal numbers of ego nodes from each of the four strata. This approach ensures sufficient statistical power for comparative analysis by preventing the smallest strata from being underrepresented and maximizing our ability to detect homophily effects in rare gender combinations, while maintaining representativeness across gender-collaboration combinations.

*Egocentric Network Extraction*: For each sampled ego, we extract their complete local network including all alters (direct collaborators) and the ties among these alters. This provides the local network structure necessary for ERGM analysis while capturing the immediate collaborative environment of each contributor.

*Post-Stratification Weighting*: All models incorporate post-stratification weights to adjust for differential sampling probabilities across strata. Each node $i$ in stratum $s$ receives a weight calculated as:

$$w_i = \frac{N_s}{n_s}$$

where $N_s$ is the total population size of stratum $s$ in the complete network and $n_s$ is the sample size for that stratum, ensuring our estimates remain representative of the broader OSS population [41].

*Analytical Applications*: The stratified egocentric networks serve as the foundation for two primary analytical approaches: (1) ERGMs to examine local network formation processes and gender homophily patterns, and (2) survival analysis to investigate how gender composition of collaborative networks influences developer retention and departure patterns.

## 3.5 Inferential Analysis: Egocentric ERGM

Our descriptive analysis answered $RQ_1$ about population-level gender homophily. However, such patterns can emerge from two distinct patterns: genuine preference for similar others (preference-driven homophily) or social structures that limit the opportunity for cross-group interaction (structurally constrained homophily) [47]. To disentangle these patterns for $RQ_1$ and answer the $RQ_2$ about patterns of tie formation, we employed egocentric Exponential-Family Random Graph Models (ERGMs).

ERGMs are statistical models that predict the probability of observing specific network structures by modeling the micro-level processes through which connections form between individuals [19, 76]. ERGMs compare the observed network against the distribution of all possible alternative networks with the same basic characteristics (*e.g.*, number of nodes), calculating how much more or less likely the observed patterns are than would occur by random chance alone. This allows ERGMs to distinguish whether observed patterns, such as people connecting preferentially to similar others, reflect genuine social preferences or simply emerge by chance given the network's composition and structural constraints, thus providing the inferential foundation for answering $RQ_2$.

Formally, an ERGM models the probability of an observed network structure through a set of sufficient statistics (network features) and their associated parameters. We estimate our models using `ergm.ego` in R [55], which implements egocentric ERGM estimation via stochastic approximation with carefully tuned MCMC (Markov Chain Monte Carlo) controls to ensure convergence [32, 55]. This approach accounts for the sampling design through post-stratification weights while avoiding the computational intractability of full-network ERGMs on million-node graphs. Common terms include: `edges` (network density baseline), `nodematch('sex', diff=TRUE)` (homophily, estimated separately for each gender category), and `nodefactor('has_women_alter')` (node-level covariates capturing whether an ego has women collaborators).

While widely applied across social sciences to study collaboration networks in organizational behavior, public health, economics, and education [28, 42], ERGMs remain underutilized in software engineering research, representing a methodological opportunity for understanding collaboration patterns in technical communities.

*3.5.1 Model Specification.* We specified a sequence of two models to formally test the patterns of tie formation. An ERGM defines the probability of an observed network, $y$, based on a set of network statistics, $g(y)$, and their corresponding parameters, $\theta$ [32]:

$$P(Y = y|\theta) = \frac{\exp(\theta^T g(y))}{c(\theta)}$$

where $c(\theta)$ is a normalizing constant. The parameters in $\theta$ represent the conditional log-odds of a tie forming. Our models analyzed tie formation among the alters within a sampled ego's local network.

*Model 1: Basic Conditional Homophily.* This model tested the most direct micro-pattern: a baseline preference for same-gender ties, controlling for local network density. The specification was:

$$\texttt{network} \sim \texttt{edges} + \texttt{nodematch('sex', diff = TRUE)} \quad (1)$$

The change in the network statistic, $\Delta g(y_{ij})$, from a tie between alters $i$ and $j$ was modeled as:

$$\begin{aligned}\Delta g(y_{ij}) = \theta_{\text{edges}} &\\ + \theta_{\text{women}} \cdot I(s_i = \text{woman} \wedge s_j = \text{woman}) &\\ + \theta_{\text{men}} \cdot I(s_i = \text{man} \wedge s_j = \text{man}) &\end{aligned} \quad (2)$$

where $I(\cdot)$ is an indicator function. A statistically significant positive coefficient for $\theta_{\text{women}}$ or $\theta_{\text{men}}$ would provide evidence for preference-driven homophily as a key micro-level pattern.

*Model 2: Collaboration Experience Effects.* This model tested a more complex pattern by examining whether an ego's own collaborative history moderates tie formation among their alters. The specification was:

$$\begin{aligned}\texttt{network} \sim \texttt{edges} + \texttt{nodematch('sex', diff = TRUE)} &\\ + \texttt{nodefactor('has\_women\_alter')} &\end{aligned} \quad (3)$$

The additional parameter, $\theta_{\text{exp}}$, for the `nodefactor` term allowed us to test whether the ego's prior exposure to women collaborators influenced the overall rate of tie formation in their local network, providing insight into how context shapes the operation of homophilous preferences.

## 3.6 Survival Analysis: Developer Retention

We define retention as continued OSS activity without an inactivity gap of at least one year, consistent with prior OSS evidence showing that return probabilities drop sharply after 12 months of inactivity [2, 4]. We operationalize retention as making at least one commit in a calendar year in any repository in the World of Code dataset. For example, a developer who commits in 2020 and 2021 is retained through 2021; if they then disappear and make no commits in 2022 or any subsequent year, they exited in 2022. A developer is considered *retained* if they are active in both 2021 and 2022 (making $\geq$1 commit in each year). Conversely, a developer *exits* if, after any active year, they do not reappear in any subsequent year up to 2022. We model time-to-exit using a weighted Cox proportional hazards model, with observations right-censored at 2022. The time origin is the developer's first active year, and the event of interest is the first year with no subsequent reappearance in the dataset.

To answer the retention component of $RQ_2$, we employed survival analysis to model factors influencing developer departure from OSS. Our observation period spanned 14 calendar years (2008-2022), with developer departure defined as not appearing in any subsequent year after being active. We analyzed developers from 2008-2021, treating those active in 2022 as right-censored.

We implemented Cox proportional hazards regression [9], which estimates the hazard rate $h(t)$ representing instantaneous departure risk:

$$h(t|\mathbf{X}) = h_0(t) \exp(\boldsymbol{\beta}^T \mathbf{X}) \quad (4)$$

where $h_0(t)$ is the baseline hazard and $\exp(\boldsymbol{\beta}^T \mathbf{X})$ captures covariate effects. The model's flexibility in not assuming a specific hazard distribution makes it suitable for the dynamic OSS ecosystem. The primary output is the Hazard Ratio (HR = $\exp(\beta_k)$), where HR < 1 indicates reduced departure risk and HR > 1 indicates increased risk. Our primary model tested:

$$h(t|\mathbf{X}) = h_0(t) \exp(\beta_1 \cdot \texttt{degree} + \beta_2 \cdot \texttt{has\_women\_alter}) \quad (5)$$

The key predictor `has_women_alter` is a binary indicator for having at least one woman collaborator in the developer's final observed year. We chose binary over continuous specification to test threshold effects and ensure statistical power across strata. We controlled for `degree` (total collaborators) to account for collaboration activity, structural opportunity, and selection effects. The variable `degree` refers to the number of unique collaborators a developer has in their final observed active year, representing their collaboration intensity at the point of potential departure. This measures the breadth of a developer's collaborative network in their last year of activity.

*Addressing Extreme Heterogeneity in Collaboration Activity*: A defining characteristic of the OSS ecosystem is extreme heterogeneity in collaboration intensity. The degree distribution in our collaboration networks is heavily right-skewed, typical of OSS and other large-scale networks [53]: many contributors link to few collaborators (median degree of 1-10), while a smaller set of core maintainers connect to hundreds or thousands of others. This creates challenges for survival analysis, as developers with vastly different activity levels may face fundamentally different departure risks.

To address this heterogeneity, we conducted robustness analyses restricting the sample to developers with degree $\geq$ 20. This

threshold serves three purposes: (1) it trims the long tail of peripheral contributors (degree <10) where cross-gender exposure is largely incidental, (2) it avoids the sparse high-degree slice (degree ≥ 30) that reduces sample size and destabilizes estimates, and (3) it ensures comparable opportunity for cross-gender ties across the analyzed population. This degree band preserves developers who are sufficiently integrated into the collaborative ecosystem for gender composition effects to be meaningful.

We conducted gender-stratified analyses rather than pooled models to accommodate different social processes: for women, ties to other women represent homophilous support connections, while for men, ties to women are heterophilous cross-gender connections. Stratification also allows baseline hazard functions to differ between genders. We tested model robustness through: (1) temporal controls using developer entry cohorts (`first_window`, the calendar year when a developer first appeared in the dataset), which account for secular trends in ecosystem evolution and cohort effects, (2) activity-stratified analysis restricting to developers with degree ≥ 20 as described above, and (3) interaction terms testing whether effects vary by collaboration volume. The proportional hazards assumption was verified using Schoenfeld residuals [25], which test whether coefficient effects remain constant over time by examining the correlation between scaled residuals and time. Under the proportional hazards assumption, Schoenfeld residuals should be independent of time; systematic patterns in residual plots or significant correlations indicate time-varying effects and violations of the assumption.

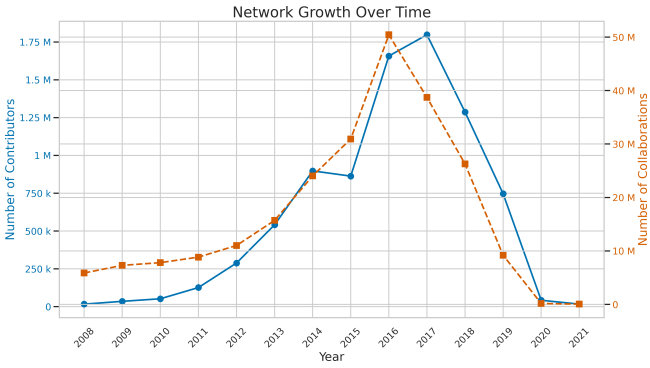
# 4 Results

## 4.1 Evidence of Gender Homophily in OSS

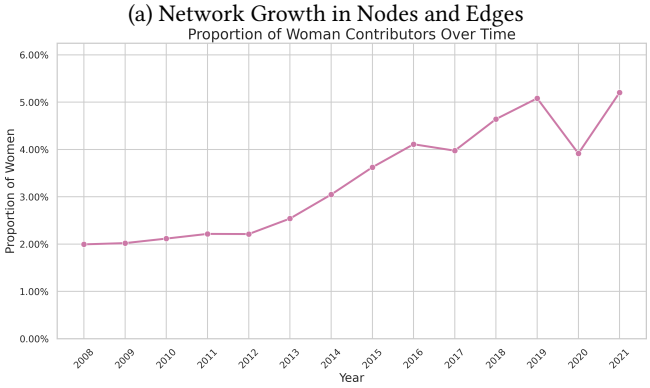
*4.1.1 Network Evolution and Composition.* We first characterize the temporal evolution and gender composition of our dataset to establish the empirical context for assessing gender homophily ($RQ_1$) and understanding the structural constraints within which tie formation occurs ($RQ_2$). Figure 2(a) illustrates the dynamic evolution of the OSS ecosystem's LCC. The LCC grew dramatically from 2008, reaching its zenith in 2016 with over 1.75 million interconnected contributors. This hyper-growth period coincides with documented surges in open source participation [21]. Following this peak, the LCC exhibits substantial contraction, with the network breaking into smaller disconnected components. This fragmentation reflects ecosystem consolidation around successful projects and natural developer turnover [18, 54, 71].

Concurrent with these structural changes, network gender composition evolved substantially (Figure 2(b)). The proportion of identifiable women contributors increased from approximately 2.0% to over 5.0% throughout our observation period, more than doubling despite the low baseline. This upward trajectory coincided with intensified focus on gender diversity within both software industry practice and academic research [73]. Our findings agreed with the "gradual but persistent" trend identified in prior OSS gender composition studies [56].

*4.1.2 Structural Proximity and Gender.* To provide initial descriptive evidence for $RQ_1$, we examine structural proximity patterns between different gender pairings. Shorter path lengths between same-gender developers would suggest clustering that exceeds what network topology alone would predict.




(a) Network Growth in Nodes and Edges




(b) Proportion of Women Contributors

**Figure 2: Temporal evolution of the OSS collaboration network from 2008 to 2022.**

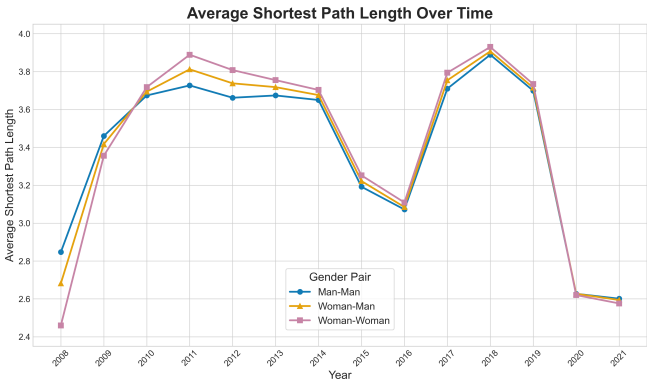



**Figure 3: Average shortest path length between gender pairs over time. Lower values indicate greater structural proximity.**

The average shortest path metric reveals dynamic evolution over the 14-year period, as shown in Figure 3. The structural proximity of different gendered dyads changes significantly over time, particularly around the network's maturity peak in 2016.

In the earlier windows (year 2008), Woman-Woman dyads exhibit greater structural proximity, with an average path length of 2.455,

compared to Woman-Man (2.680) and Man-Man (2.846) pairs. This pattern suggests relatively high initial cohesion within the female developer community. However, this pattern inverts during the network's high-growth phase (2010-2016). For instance, in 2011, the average path length for the Woman-Woman pairs (3.889) becomes the longest, while the Man-Man pairs (3.728) become the most structurally proximate. This trend, where same-gender women's ties are more distant than other pairs, persists for several years, suggesting that as the ecosystem rapidly expanded, the structural distance between women increased.

Finally, in the most recent years of our observation (2020 and 2021), path lengths for all three groups converged, decreasing to their lowest levels. This final shift may reflect a more mature and integrated state of the ecosystem. The dynamic nature of these path lengths across different eras highlights that the role and position of gender in network structure are not fixed, but evolve in response to the overall growth and changes within the OSS community.

In summary, structural proximity analysis reveals that gender-based clustering patterns are neither static nor uniform across time, instead reflecting the broader developmental trajectory of the OSS ecosystem itself.

*4.1.3 Quantifying Gender Homophily.* To directly address $RQ_1$, we examined O/E ratios for different gender pairings. Figure 4 presents these ratios for direct and indirect connections. For direct collaborations (path length = 1, Figure 4(a)), we found strong evidence of homophily among women. The O/E ratio for Woman-Woman ties is consistently and substantially greater than 1.0, peaking at over 2.2. This indicates that women collaborate directly with other women at more than double the rate expected by chance. The ratio for Man-Man ties remains stable and slightly above 1.0, indicating a mild but persistent preference for same-gender collaboration among men. Correspondingly, the ratio for Woman-Man ties is consistently below 1.0, suggesting that cross-gender collaborations are less frequent than expected.

This pattern evolves at longer distances. At a path length of 2 (Figure 4(b)), the O/E ratio for Woman-Woman pairs drops significantly, often falling below 1.0. This suggests that while women form tight, direct-tie clusters, these clusters may be structurally segregated from one another within the broader network, making two-step paths between women less common than expected. The patterns at path length 3 (Figure 4(c)) show a convergence toward 1.0 for all groups, indicating that at greater distances, the effects of local homophily dissipate as connections become more random.

Collectively, these descriptive findings establish that gender is a significant organizing principle in the OSS collaboration network. The observed patterns of structural proximity and tie formation provide a strong empirical basis for our subsequent inferential analyses into the patterns and consequences of this gender-based sorting.

## 4.2 ERGM Results: Evolving patterns of Gender-Based Tie Formation

To address $RQ_2$, we employed egocentric ERGMs to identify micro-level processes driving collaboration while controlling for structural factors. Our analysis reveals that while gender homophily is significant, its effects are deeply intertwined with the evolving influence of a developer's collaborative context. Specifically, we find that women consistently show a preference for same-gender ties, while men show a preference for cross-gender ties. However, the effect of having at least one woman collaborator (`has_women_alter`) is a more powerful predictor of local network activity, and the function of this pattern fundamentally changes as the ecosystem matures.

*4.2.1 Quantifying the Core patterns.* Table 1 presents ERGM results for two representative years (2012 and 2019) that illustrate the temporal evolution of tie formation patterns. Across the entire 14-year period, our models identify three consistent, significant patterns. First, for women's homophily (the `nodematch.sex.1` term), the average significant coefficient was positive ($\beta = 0.375$). This translates to an odds ratio of 1.46, indicating that, on average, a tie between two women was 1.46 times more likely to form than a tie between a woman and a man, holding other factors constant.

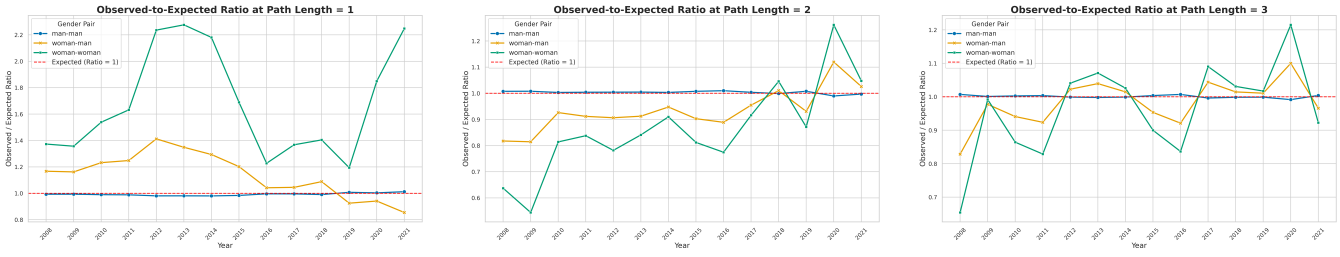
Second, for men's homophily, the effect was consistently negative ($\beta = -1.212$). This yields an odds ratio of 0.30, meaning a tie between two men was only 0.30 times as likely to form as a tie between a man and a woman. This reveals a strong and persistent pattern of cross-gender preference for men at the local network level.

Third, the most powerful pattern was the effect of having prior experience with women collaborators. Across all significant results, this term had an average coefficient of 1.141, corresponding to an odds ratio of 3.13. This means that, on average, the presence of at least one woman collaborator in an ego's network was associated with a 3.13-fold increase in the likelihood of their collaborators forming ties with each other.

*4.2.2 Temporal Dynamics: From Experience Boost to Structural Signal.* While the overall averages are informative, Figure 5, showing the basic homophily model, displays considerable year-to-year volatility in the gender preference terms. This instability suggests that Model 1, by itself, provides an incomplete picture.

This picture is clarified by Model 2 (Table 1), which adds the `has_women_alter` term (Figure 6). The introduction of this control variable stabilizes the homophily estimates, particularly by attenuating the extreme spikes observed around 2014. This indicates that much of the yearly fluctuation in simple gender preference was, in fact, driven by the underlying effect of being in a gender-diverse collaborative context.

The year-to-year volatility in gender preference coefficients, particularly the sharp changes observed around 2013-2014, likely reflects multiple converging factors in the OSS ecosystem's evolution. This period coincided with GitHub's transition from primarily patch-based to pull-request-centered workflows, which fundamentally altered how developers discovered collaborators and formed cross-project connections [24, 36, 57, 77]. The shift to pull requests as the dominant contribution mechanism expanded the visibility of contributions across projects, potentially changing the structural opportunities for gender-based tie formation. Additionally, this era saw rapid platform growth with influx of new contributors, evolving code review norms and tooling adoption, and changes in repository metadata coverage in the World of Code dataset. These contextual factors emphasize that homophily operates within, and is shaped by, the broader sociotechnical infrastructure of software




(a) Path Length = 1 (Direct Ties) (b) Path Length = 2 (c) Path Length = 3

**Figure 4: Observed-to-Expected (O/E) Ratios for dyad connections at different path lengths. A ratio > 1.0 indicates more connections than expected by chance. (a) At path length 1, strong homophily is evident. (b, c) At longer path lengths, Woman-Woman pairs become less frequent than expected, suggesting local clustering but broader network segregation.**

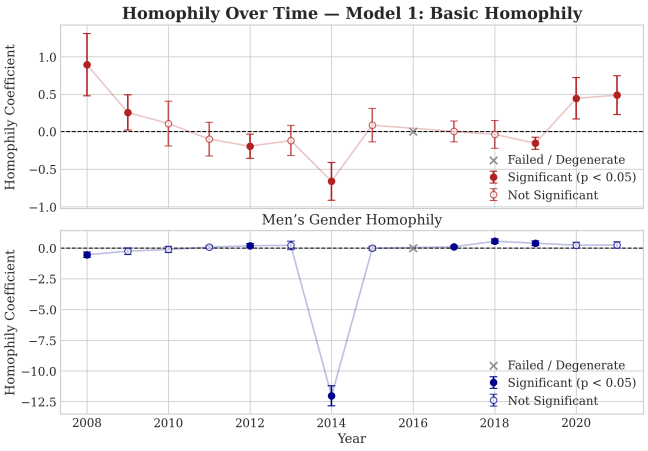



**Figure 5: Homophily Over Time (Model 1: Basic Homophily). The plots show the coefficient for women's (top) and men's (bottom) same-gender tie preference for each year.**

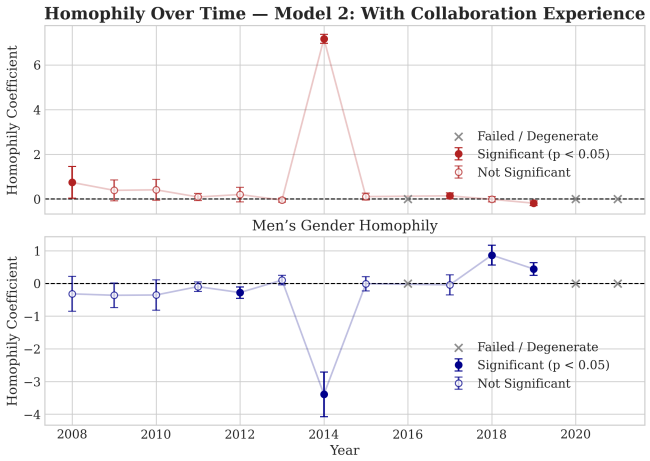



**Figure 6: Homophily Over Time (Model 2: With Collaboration Experience).**

development platforms. We offer this as historical context rather than a causal claim about specific mechanisms.

Table 1 provides a direct comparison that highlights the evolution of this pattern. In 2012, the coefficient for `has_women_alter` was large and positive ($\beta = 2.58$), indicating that being in a network with women collaborators was associated with a massive increase in local tie formation. By 2019, however, the coefficient had inverted, becoming small but significantly negative ($\beta = -0.10$). This inversion is a key finding: it suggests that the social function of having women collaborators changed over time. What began as a pattern associated with a general increase in network activity may have, in a more mature and dense ecosystem, become a signal of a specific type of structural position with different collaborative properties.

*4.2.3 A Temporal Story of Tie Formation.* Our egocentric ERGM analysis provides a multi-faceted answer to $RQ_2$ by identifying the patterns of tie formation and charting their evolution. We find that:

**Gendered preference is a persistent pattern**, but it operates differently for men and women. The tendency for women to form same-gender ties (`nodematch.sex.1`) and for men to form

**Table 1: Egocentric ERGM Results for 2012 and 2019.**

| Year | Term | Model 1 | Model 2 |
|---|---|---|---|
| **2012** | edges | 3.84*** (0.07) | -0.64 (0.68) |
| | nodematch.sex.1 | -0.19* (0.08) | 0.20 (0.17) |
| | nodematch.sex.-1 | 0.20* (0.09) | -0.28** (0.09) |
| | nodefactor.has_women_alter.1 | – | 2.58*** (0.35) |
| **2019** | edges | 2.86*** (0.03) | 3.01*** (0.05) |
| | nodematch.sex.1 | -0.15*** (0.04) | -0.19*** (0.05) |
| | nodematch.sex.-1 | 0.41*** (0.10) | 0.44*** (0.10) |
| | nodefactor.has_women_alter.1 | – | -0.10* (0.04) |

* p<0.05, ** p<0.01, *** p<0.001.
The table shows coefficients with standard errors in parentheses.

cross-gender ties (`nodematch.sex.-1`) are stable features of the ecosystem's micro-dynamics.

**Collaborative context is a powerful, evolving pattern.** The effect of having at least one woman collaborator (`nodefactor.has_women_alter.1`) is not constant. Its shift from being a strong positive predictor of tie formation in the early years to a negative predictor in later years is direct evidence that the patterns of collaboration change as the social structure of the ecosystem matures. These findings suggest a process of network evolution

**Table 2: Descriptive Statistics by Developer Group**

| Group (at final observation) | N | % Departed | Mean Degree | Median Degree |
|---|---|---|---|---|
| Women w/o women collaborators | 9,125 | 100% | 3.2 | 1.0 |
| Women w/ women collaborators | 9,421 | 98.8% | 65.7 | 10.0 |
| Men w/o women collaborators | 43,662 | 100% | 4.1 | 1.0 |
| Men w/ women collaborators | 48,409 | 99.4% | 277.6 | 33.0 |

**Table 3: Cox Regression Results on Developer Departure**

| Gender | Hazard Ratio (HR) | 95% Confidence Interval | p-value |
|---|---|---|---|
| Women | 0.974 | (0.940, 1.009) | 0.138 |
| Men | 0.988 | (0.975, 1.002) | 0.100 |

where the impact of individual attributes (like being in a gender-diverse network) changes over time. Initially, it may facilitate general network growth, but as the network solidifies, these same attributes become signals of an individual's embedded structural position. This temporal shift aligns with dynamic network theories where initial, preference-based choices crystallize into enduring structures that then constrain future behavior [58].

## 4.3 Retention Analysis Results: Associations Between Gender-Diverse Collaboration and Retention Patterns

To address the second part of $RQ_2$, how gender-based patterns correlate with retention outcomes, we examined developer departure risk using Cox proportional hazards models. Our analysis reveals that the presence of women collaborators constitutes a significant protective effect against developer departure. This effect benefits both women and men, but emerges only after controlling for a developer's total collaboration activity. The aggregate analysis initially obscures this relationship due to severe confounding by collaboration intensity.

The analysis was based on a final sample of 111,773 unique developers from the stratified egocentric networks, observed over 14 annual time windows (2008-2022). After cleaning for model inclusion, the sample consisted of 18,546 women (16.8%) and 92,071 men (83.2%). The event of interest, developer departure, was observed for 99.6% of these developers, and this near-universal departure confirms the high-turnover nature of the ecosystem.

A defining characteristic of this ecosystem is extreme heterogeneity in collaboration intensity, measured by network degree (unique collaborators in a developer's final active year). The developer population spanned from casual, peripheral contributors to highly-connected core maintainers, documented in Table 2.

*4.3.1 The Confounding Effect of Collaboration Activity.* We first estimated our primary Cox proportional hazards model to assess the effect of having at least one woman collaborator on the hazard of departure, while statistically controlling for degree. This initial model yielded non-significant results (Table 3).

However, the diagnostic statistics in Table 2 revealed that this result was an artifact of severe confounding. Men with at least one female collaborator had, on average, 67 times more collaborators than men without. This indicated that the primary model was not comparing similar individuals but rather comparing peripheral, casual contributors to highly-central core developers.

**Table 4: Confounding-Controlled Cox Regression Results**

| Model Specification | Gender | HR | 95% CI | Hazard Reduction |
|---|---|---|---|---|
| **Cohort-Controlled** | Women | 0.933*** | (0.908, 0.958) | 6.7% |
| | Men | 0.983** | (0.971, 0.996) | 1.7% |
| **Activity-Matched** | Women | 0.793** | (0.665, 0.945) | 20.7% |
| (Degree $\geq$ 20) | Men | 0.899*** | (0.851, 0.949) | 10.1% |

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

*4.3.2 Revealing the Association Through Stratified Analysis.* To isolate the effect of woman collaborator presence, we implemented two refined model specifications designed to address the identified confounding. First, we incorporated temporal controls by including developer entry cohort (`first_window`) to account for secular trends in ecosystem evolution. Second, we conducted activity-stratified analysis restricting the sample to developers with degree $\geq$ 20, ensuring comparison among similarly engaged community members. Table 4 demonstrates that proper confounding control unveils statistically significant protective effects of woman collaborator presence across both gender groups. The cohort-controlled specification indicates 6.7% departure hazard reduction for women (HR=0.933, p<0.001) and 1.7% reduction for men (HR=0.983, p=0.01). The activity-matched analysis yields more pronounced effects: among active developers (degree $\geq$ 20), woman collaborator presence corresponds to 20.7% hazard reduction for women (HR=0.793, p=0.009) and 10.1% reduction for men (HR=0.899, p<0.001).

These findings demonstrate the critical importance of controlling for collaboration activity in open-source retention analyses. The initial aggregate analysis was misleading due to systematic differences in collaboration intensity between comparison groups. Through activity stratification, we reveal the true relationship: among developers sufficiently integrated into the collaborative ecosystem, ties to woman collaborators are associated with significant retention improvements. These associations appear consistent across both gender groups, indicating that gender-diverse collaboration benefits extend beyond simple homophilous matching to represent a fundamental characteristic of well-integrated, diverse network structures. The magnitude of these effects, particularly the 20.7% hazard reduction for active women, represents practically significant improvements in developer retention, with clear implications for open source project sustainability and diversity initiatives.

# 5 Discussion

Our longitudinal analysis of 1.6 million OSS developers reveals how gender shapes collaboration networks and retention patterns. We discuss the theoretical and practical implications of these findings.

## 5.1 Asymmetric Homophily Reflects Minority Group Adaptation

Addressing $RQ_1$, our findings reveal not only that homophily exists, but that it operates asymmetrically across gender groups. This asymmetry challenges standard assumptions in network analysis that homophily operates symmetrically [47]. For women comprising 2-5% of the network, same-gender ties likely provide critical support

and authentic participation spaces that majority-group men do not need [33, 38]. Our findings extend prior OSS gender research [17, 68] by revealing the network structures women construct as adaptive strategies in male-dominated environments, rather than merely documenting barriers to participation. The temporal dynamics mirrors the sociological findings that minority cohesion fluctuates with environmental pressures [54]. This suggests gender homophily is strategic adaptation rather than merely individual preference, with important implications for understanding how underrepresented groups navigate technical communities.

## 5.2 Collaborative Context Matters More Than Individual Preferences

Addressing the first part of $RQ_2$, our ERGM analyses reveal that collaborative context shapes network evolution more powerfully than individual gender preferences alone. The finding that having women collaborators predicts tie formation more strongly than gender homophily, and that this effect reverses over time, suggests network evolution where preference-based choices crystallize into structural positions whose meanings shift [58, 65]. This temporal shift has theoretical implications: in mature networks, gender-diverse ties may signal membership in specific sub-communities rather than general openness to diversity. This interpretation aligns with research on structural embeddedness, where the *meaning* of ties changes as networks evolve from fluid to consolidated [26].

## 5.3 Retention Effects Highlight Diversity as Infrastructure, Not Decoration

Addressing the second part of $RQ_2$, our survival analysis reveals that gender-diverse collaboration significantly reduces departure risk after controlling for confounding by collaboration intensity. This demonstrates that diversity initiatives address tangible structural disadvantages, not merely symbolic goals. The protective effect emerges only through careful stratification, highlighting a critical methodological insight: aggregate analyses of diversity effects can be misleading without accounting for systematic differences in network positioning. This finding complements prior OSS retention research on technical factors [79] and early experiences [66] by revealing social network composition as an independent retention mechanism. The larger benefits for women align with organizational research showing diverse networks disproportionately help minority members [34].

## 5.4 Implications for OSS Sustainability and Software Engineering Practice

Together, these findings have several important implications. First, they suggest that fostering gender-diverse collaboration can serve as an effective strategy to improve contributor retention and project sustainability. This aligns with calls from prior work advocating for increased inclusivity to enhance OSS community health [15, 51]. Second, the observed clustering and potential fragmentation of women's networks during growth phases raise concerns about isolated sub-communities that may hinder cross-group information flow and opportunities. This echoes sociological insights about the risks of homophilous segregation limiting access to broader resources and influence [47, 58]. Thus, OSS projects and platform designers should consider mechanisms to bridge gender clusters and foster integrative ties, potentially through targeted community-building or recommendation systems [35]. Third, the temporal variation in gender homophily's effects implies that diversity and inclusion efforts should be adapted to the maturity and growth stage of OSS projects, reflecting the dynamic social processes shaping collaboration [30].

## 5.5 Limitations and Future Directions

Our study has several limitations that suggest directions for future work. First, causal inference requires experimental or quasi-experimental designs exploiting natural experiments in OSS platforms. Second, name-based gender inference introduces measurement error, with 30.93% unknown classifications potentially biasing results toward conventional naming patterns and excluding developers using pseudonyms or non-binary identities [37]. Future work should leverage self-reported gender data when available and develop methods for analyzing non-binary gender categories. Third, unobserved confounding from project quality, developer expertise, or institutional affiliations could drive both network composition and retention. Addressing this requires individual-level controls and project fixed-effects models. Fourth, our binary gender classification obscures within-group heterogeneity across race, nationality, and other dimensions [10]. Finally, qualitative research is needed to identify the mechanisms—mentorship quality, psychological safety, or knowledge access—underlying the protective effect of gender-diverse collaboration [17, 69].

# 6 Conclusion

Gender significantly shapes collaboration patterns in OSS development, with homophilous networks correlating with improved developer retention. This suggests that the composition of the network may be an important factor in the sustainability of the software project. As OSS development relies on distributed, voluntary collaboration, understanding the social dynamics that sustain developer communities becomes essential. Future research should investigate causal mechanisms and develop network-aware interventions to build more inclusive and resilient software development ecosystems.

# Acknowledgments

This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology.

# References

[1] Dominic Abrams and Michael A Hogg. 1999. *Social identity and social cognition.* Blackwell Oxford.

[2] Guilherme Avelino, Eleni Constantinou, Marco Tulio Valente, and Alexander Serebrenik. 2019. On the abandonment and survival of open source projects: An empirical investigation. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM).* IEEE, 1–12.

[3] Donn Byrne. 1997. An overview (and underview) of research and theory within the attraction paradigm. *Journal of Social and Personal Relationships* 14, 3 (1997), 417–431.

[4] Fabio Calefato, Marco A. Gerosa, Giuseppe Iaffaldano, Filippo Lanubile, and Igor Steinmacher. 2022. Will You Come Back to Contribute? Investigating the Inactivity of OSS Core Developers in GitHub. *Empirical Software Engineering* 27, 3 (2022), 76. doi:10.1007/s10664-021-10012-6

[5] Elian Carsenat. 2019. Inferring gender from names in any region, language, or alphabet. doi:10.13140/RG.2.2.11516.90247

[6] Gemma Catolino, Fabio Palomba, Damian A Tamburri, Alexander Serebrenik, and Filomena Ferrucci. 2019. Gender diversity and women in software teams: How do they affect community smells?. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, 11–20. doi:10.1109/ICSE-SEIS.2019.00010

[7] Jailton Coelho and Marco Tulio Valente. 2017. Why modern open source projects fail. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*. 186–196.

[8] E. Cohen and M. Consens. 2018. Large-Scale Analysis of the Co-Commit Patterns of the Active Developer Network. In *Proceedings of the 15th International Conference on Mining Software Repositories (MSR)*. https://www.cs.toronto.edu/~consens/AnalysisGitHubCoCommit/GitHubCoCommitAnalysisCohenConsensMSR2018.pdf

[9] David R Cox. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* 34, 2 (1972), 187–220.

[10] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1989, 1 (1989), 139–167.

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

[12] Laura Dabbish, Colleen Stuart, Jason Tsay, and James Herbsleb. 2012. Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository. In *Proceedings of CSCW*. 1277–1286. doi:10.1145/2145204.2145396

[13] Adriana Domínguez-Díaz, Vanessa Terán-Messa, Valentina Vargas-Calderón, and Miguel Gómez-García. 2024. Comparative analysis of automatic gender detection from names: evaluating the stability and performance of ChatGPT versus NamSor, and Gender-API. *PeerJ Computer Science* 10 (2024), e2378. doi:10.7717/peerj-cs.2378

[14] Nadia Eghbal. 2016. *Roads and Bridges: The Unseen Labor Behind Our Digital Infrastructure.* Technical Report. Ford Foundation. https://www.fordfoundation.org/work/learning/research-reports/roads-and-bridges-the-unseen-labor-behind-our-digital-infrastructure/

[15] Yulin Fang and Derrick Neufeld. 2009. Understanding sustained participation in open source software projects. *Journal of Management Information Systems* 25, 4 (2009), 9–50.

[16] Marcos Oliveira Fariba Karimi. 2023. On the inadequacy of nominal assortativity for assessing homophily in networks. *Scientific Reports* (2023).

[17] Denae Ford, Justin Smith, Philip J. Guo, and Chris Parnin. 2016. Paradise Unplugged: Identifying Barriers for Female Participation on Stack Overflow. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2016)*. ACM, New York, NY, USA, 846–857. doi:10.1145/2950290.2950331

[18] Matthieu Foucault, Mickaël Palyart, Xavier Blanc, Gail C. Murphy, and Jean-Rémy Falleri. 2015. Impact of Developer Turnover on Quality in Open-Source Software. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2015)*. 829–841.

[19] Ove Frank and David Strauss. 1986. Markov graphs. *J. Amer. Statist. Assoc.* 81, 395 (1986), 832–842.

[20] Min Gao, Zheng Chen, Ruichen Yao, Chenhao Li, Xin Wang, Yupeng Zhao, Wenjie Wang, Hui Zhang, Jingjing Gao, Yaqi Ding, et al. 2023. EasyGraph: A multifunctional, cross-platform, and effective library for interdisciplinary network analysis. *Patterns* 4, 10 (2023), 100835.

[21] GitHub, Inc. 2017. The State of the Octoverse 2017. https://octoverse.github.com/2017/.

[22] Steven M Goodreau, James A Kitts, and Martina Morris. 2009. Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* 46, 1 (2009), 103–125.

[23] Christoph Gote, Ingo Scholtes, and Frank Schweitzer. 2021. Analysing time-stamped co-editing networks in software development teams using git2net. *Empirical Software Engineering* 26, 75 (2021). doi:10.1007/s10664-020-09928-2

[24] Georgios Gousios, Martin Pinzger, and Arie van Deursen. 2014. An Exploratory Study of the Pull-Based Software Development Model. In *Proceedings of MSR*. 345–355. doi:10.1145/2597073.2597074

[25] Patricia M. Grambsch and Terry M. Therneau. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81, 3 (09 1994), 515–526. doi:10.1093/biomet/81.3.515 arXiv:https://academic.oup.com/biomet/article-pdf/81/3/515/714158/81-3-515.pdf

[26] Mark Granovetter. 1985. Economic action and social structure: The problem of embeddedness. *Amer. J. Sociology* 91, 3 (1985), 481–510.

[27] Jungpil Hahn, Jae Yoon Moon, and Chen Zhang. 2006. Impact of social ties on open source project team formation. In *IFIP international conference on open source systems*. Springer, 307–317.

[28] Jenine K Harris et al. 2023. The application of exponential random graph models to collaboration networks in biomedical and health sciences: a review. *Network Modeling Analysis in Health Informatics and Bioinformatics* 12 (2023).

[29] Pamela J Hinds, Kathleen M Carley, David Krackhardt, and Doug Wholey. 2000. Choosing work group members: Balancing similarity, competence, and familiarity. *Organizational behavior and human decision processes* 81, 2 (2000), 226–251.

[30] Petter Holme and Jari Saramäki. 2012. Temporal networks. *Physics Reports* 519, 3 (2012), 97–125. doi:10.1016/j.physrep.2012.03.001

[31] Yun Huang, Cuihua Shen, Dmitri Williams, and Noshir Contractor. 2009. Virtually there: Exploring proximity and homophily in a virtual world. In *2009 International Conference on Computational Science and Engineering*, Vol. 4. IEEE, 354–359.

[32] David R Hunter, Mark S Handcock, Carter T Butts, Steven M Goodreau, and Martina Morris. 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software* 24, 3 (2008), 1–29.

[33] Herminia Ibarra. 1992. Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative science quarterly* 37, 3 (1992), 422–447.

[34] Herminia Ibarra. 1993. Personal networks of women and minorities in management: A conceptual framework. *Academy of Management Review* 18, 1 (1993), 56–87.

[35] Mitchell Joblin, Wolfgang Mauerer, Sven Apel, Janet Siegmund, and Dirk Riehle. 2015. From developer networks to verified communities: A fine-grained approach. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. IEEE, 563–573.

[36] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M. German, and Daniela Damian. 2014. The Promises and Perils of Mining GitHub. In *Proceedings of the 11th Working Conference on Mining Software Repositories (MSR)*. https://kblincoe.github.io/publications/2014_MSR_Promises_Perils.pdf

[37] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[38] Adam M Kleinbaum, Toby E Stuart, and Michael L Tushman. 2013. Discretion within constraint: Homophily and structure in a formal organization. *Organization Science* 24, 5 (2013), 1316–1336.

[39] Gueorgi Kossinets and Duncan J Watts. 2009. Origins of homophily in an evolving social network. *Amer. J. Sociology* 115, 2 (2009), 405–450. doi:10.1086/599247

[40] Pavel N Krivitsky. 2012. Exponential-family random graph models for valued networks. *Electronic journal of statistics* 6 (2012), 1100.

[41] Pavel N Krivitsky, Martina Morris, and Michał Bojanowski. 2022. Impact of survey design on estimation of exponential-family random graph models from egocentrically-sampled data. *Social Networks* 69 (2022), 22–34. doi:10.1016/j.socnet.2020.10.001

[42] Dean Lusher, Johan Koskinen, and Garry Robins. 2013. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

[43] Yuxing Ma, Tapajit Dey, Chris Bogart, Sadika Amreen, Marat Valiev, Adam Tutko, David Kennard, Russell Zaretzki, and Audris Mockus. 2021. World of code: enabling a research workflow for mining and analyzing the universe of open source VCS data. *Empirical Software Engineering* 26 (2021), 1–42.

[44] Eleanor E Maccoby. 1990. Gender and relationships: A developmental account. *American Psychologist* 45, 4 (1990), 513–520.

[45] Eleanor E Maccoby. 1998. Gender segregation in the workplace. In *The Developmental Social Psychology of Gender*. Psychology Press, 365–398.

[46] Eleanor E Maccoby. 1998. *The Two Sexes: Growing Up Apart, Coming Together*. Harvard University Press, Cambridge, MA.

[47] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[48] Clare M Mehta and JoNell Strough. 2009. Sex segregation in friendships and normative contexts across the life span. *Developmental review* 29, 3 (2009), 201–220.

[49] 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* 11, 3 (2002), 309–346. doi:10.1145/567793.567795

[50] Peter R Monge, Noshir S Contractor, Peter S Contractor, R Peter, S Noshir, et al. 2003. *Theories of communication networks*. Oxford University Press, USA.

[51] Dawn Nafus. 2012. 'Patches don't have gender': What is not open in open source software. *New Media & Society* 14, 4 (2012), 669–683. doi:10.1177/1461444811422887

[52] Frank Nagle, Manuel Hoffmann, and Yanuo Zhou. 2024. The $8.8 trillion value of open source software. *Harvard Business School Working Paper* (2024). Available at: https://www.hbs.edu/faculty/Pages/item.aspx?num=65230.

[53] M. E. J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Rev.* 45, 2 (2003), 167–256. doi:10.1137/S003614450342480

[54] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. 2007. Quantifying social group evolution. *Nature* 446, 7136 (2007), 664–667.

[55] Martina Morris Pavel N. Krivitsky. 2017. Inference for social network models from egocentrically sampled data, with application to understanding persistent

racial disparities in HIV prevalence in the US. *Ann. Appl. Stat.* (2017).

[56] Huilian Sophie Qiu, Zihe H Zhao, Justin Wang, Tielin Katy Yu, Alexander Ma, Hongbo Fang, Laura Dabbish, and Bogdan Vasilescu. 2023. Gender Representation Among Contributors to Open-Source Infrastructure - An Analysis of 20 Package Manager Ecosystems. In *International Conference on Software Engineering, Software Engineering in Society (ICSE SEIS)*. ACM.

[57] Foyzur Rahman and Chanchal K. Roy. 2014. An Insight into the Pull Requests of GitHub. In *Proceedings of ICSE Companion*. 384–387. doi:10.1145/2591062.2591126

[58] Mark T Rivera, Sara B Soderstrom, and Brian Uzzi. 2010. Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *annual Review of Sociology* 36 (2010), 91–115.

[59] Gregorio Robles, Laura Arjona Reina, Jesus M Gonzalez-Barahona, and Santiago Duenas Dominguez. 2016. Women in free/libre/open source software: The situation in the 2010s. In *IFIP International Conference on Open Source Systems*. Springer, 163–173. doi:10.1007/978-3-319-39225-7_13

[60] Martin Ruef, Howard E Aldrich, and Nancy M Carter. 2003. The structure of founding teams: Homophily, strong ties, and isolation among US entrepreneurs. *American sociological review* (2003), 195–222.

[61] Lucía Santamaría and Helena Mihaljević. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* 4 (2018), e156. doi:10.7717/peerj-cs.156

[62] Paul Sebo. 2021. Performance of gender detection tools: a comparative study of name-to-gender inference services. *Journal of the Medical Library Association* 109, 3 (2021), 414–421. doi:10.5195/jmla.2021.1208

[63] Paul Sebo. 2022. How accurate are gender detection tools in predicting the gender for Chinese names? *Journal of the Medical Library Association* 110, 2 (2022), 205–211. doi:10.5195/jmla.2022.1396

[64] Leif Singer, Fernando Figueira Filho, Brendan Cleary, Christoph Treude, Margaret-Anne Storey, and Kurt Schneider. 2014. Software Engineering at the Speed of Light: How Developers Stay Current Using Twitter. In *Companion Proceedings of ICSE*. 211–221. doi:10.1145/2568225.2568305

[65] Christian Steglich, Tom A B Snijders, and Michael Pearson. 2010. Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology* 40, 1 (2010), 329–393. doi:10.1111/j.1467-9531.2010.01225.x

[66] Igor Steinmacher, Marco Aurélio Graciotto Silva, Marco Aurélio Gerosa, and David F Redmiles. 2015. Social barriers faced by newcomers placing their first contribution in open source software projects. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1379–1392.

[67] Sayma Sultana, Zadia Codabux Turzo, and Amiangshu Bosu. 2023. Code reviews in open source projects: how do gender biases affect participation and outcomes? *Empirical Software Engineering* 28, 4 (2023), 92.

[68] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, and Chris Parnin. 2016. Gender bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science* (02 2016). doi:10.7287/PEERJ.PREPRINTS.1733V1

[69] Bianca Trinkenreich, Igor Wiese, Anita Sarma, Marco Gerosa, and Igor Steinmacher. 2021. Hidden figures: Roles and pathways of successful OSS contributors. In *Proceedings of the ACM/IEEE 43rd International Conference on Software Engineering*. 1347–1358.

[70] John C Turner, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell. 1987. *Rediscovering the social group: A self-categorization theory.* basil Blackwell.

[71] Marat Valiev, Bogdan Vasilescu, and James D. Herbsleb. 2018. Ecosystem-level Determinants of Sustained Activity in Open-Source Projects: A Case Study of the PyPI Ecosystem. In *Proceedings of the 2018 26th ACM Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2018)*. 644–655.

[72] Johannes van der Pol. 2019. Introduction to Network Modeling Using Exponential Random Graph Models (ERGM): Theory and an Application Using R-Project. *Computational Economics* 54 (2019), 845–875.

[73] Bogdan Vasilescu, Dror Posnett, Baishakhi Ray, Mark G. J. 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 (CHI '15)*. ACM, New York, NY, USA, 3789–3798. doi:10.1145/2702123.2702549

[74] Balázs Védres and Orsolya Vásárhelyi. 2019. Gendered behavior as a disadvantage in open source software development. *EPJ Data Science* 8, 1 (2019), 25. doi:10.1140/epjds/s13688-019-0202-z

[75] Lois M Verbrugge. 1977. The structure of adult friendship choices. *Social forces* 56, 2 (1977), 576–597.

[76] Stanley Wasserman and Philippa Pattison. 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika* 61, 3 (1996), 401–425.

[77] Wired. 2013. From Collaborative Coding to Wedding Invitations: GitHub Is Going Mainstream. https://www.wired.com/2013/09/github-for-anything. Accessed November 2025.

[78] B Wu and C Wu. 2021. Research on the mechanism of knowledge diffusion in the MOOC learning forum using ERGMs. *Computers & Education* 173 (2021), 104295.

[79] Shurui Zhou and Audris Mockus. 2012. What makes a good commit message?. In *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*. 1–11.