The Structure of Cross-National Collaboration in Open-Source Software Development

Henry Xu
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
henryxu@mit.edu

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

Hao He
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
haohe@andrew.cmu.edu

Hongbo Fang
University of Chicago
Chicago, Illinois, USA
fanghongdoublebo@gmail.com

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

Patrick S. Park
Carnegie Mellon University
Pittsburgh, USA
patpark@cmu.edu

Abstract

Open-source software (OSS) development platforms, such as GitHub, expand the potential for cross-national collaboration among developers by lowering the geographic, temporal, and coordination barriers that limited software innovation in the past. However, research has shown that the technological affordances that facilitate cross-national collaboration do not uniformly benefit all countries. Using the GitHub Innovation Graph dataset, which aggregates the complete cross-country collaborations among the entire population of GitHub developers, we present quantitative evidence of deep-seated religious and cultural affinities, shared colonial histories, and geopolitical factors structuring the collaborations between non-U.S. country pairs that become visible when the overarching dominance of the U.S. is removed from the data. This study highlights the opportunities to develop decentralizing strategies to facilitate new collaborations between developers in non-U.S. countries, thereby fostering the development of novel, innovative solutions. More generally, this study also underscores the importance of contextualizing user behavior and knowledge management in information systems with long-term, macro-social conditions in which these systems are inextricably embedded.

Keywords

open-source software, international collaboration, homophily, network analysis, GitHub, exponential random graph models, block models, world systems theory, colonialism, geopolitics

ACM Reference Format:
Henry Xu, Katy Yu, Hao He, Hongbo Fang, Bogdan Vasilescu, and Patrick S. Park. 2025. The Structure of Cross-National Collaboration in Open-Source Software Development. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746252.3761237

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761237

1 Introduction

How do countries collaborate in open-source software development? Throughout history, dense economic trade relations in commodities, goods, services, and labor tended to promote international peace whereas sparse connections and fragmentation heightened the probability of war [13]. These international connections among countries have been subject to a myriad of social, economic, and political forces with shifting influences over time. For example, Western capitalist versus Eastern communist countries during the Cold War era formed connections along ideological fault lines while the post-Cold War divisions have been structured in part by deep-seated culture and religion [12]. Their significance has been shown to persist in online social interactions among millions of email and social media users, despite the friction-less connectivity that these communication technologies enabled [24].

Although studies show the persistence of such ideological, cultural, and geopolitical forces on cross-national online communication, it is an open question whether cross-national collaborations with a clearer instrumental focus, such as those observed in open-source software development, also form under the influence of such macro-social forces. Presumably, these bottom-up collaborations between millions of software developers across national borders should appear oblivious to these social forces insofar as the online collaboration platform helps overcome spatio-temporal barriers of collaboration in the creation of technological and economic value. Indeed, previous studies of open-source software development have shown that cross-country collaborations largely exhibit such instrumental rationality, reflecting the economic dominance of the U.S. and, to a lesser extent, of a handful of technologically advanced European countries. Although this heavy reliance on the U.S. is taken for granted, it is puzzling from the vantage point of the potential for frictionless and untethered collaborations that OSS platforms afford.

Based on a recent cross-country collaboration dataset released by GitHub, we construct a network of OSS collaborations between countries and study the patterns hidden under the global dominance of the United States. We reveal the structural markers of international OSS collaboration indicative of hierarchical organization in tandem with deep-seated, macro-cultural fault lines that have been observed in cross-border email and Twitter communications [24]. Despite the common belief that OSS collaboration can overcome geographical barriers [4], we observe cultural affinities (i.e., cultural homophily) and historical influences (e.g., shared colonial past) that create social boundaries in inter-country collaborations.

Our paper makes two primary contributions. First, we contribute to OSS research by proposing to remove the dominant U.S. from analysis when exploring otherwise hidden structures that provide insight. Second, we contribute to data mining and knowledge management in collaborative information systems by demonstrating the deep insights that can be gained from considering social processes that unfold at historical time scales, an approach that has been neglected in the literature.

2 Related Work

2.1 Structure of Global Interdependence

A prevailing view on international relations is that the interdependence among nations lead to a more integrated world. Dense connections between countries through economic exchange of goods, services, and labor create strong economic interdependence that inhibit war and lead to prolonged prosperity [13]. From this perspective, the OSS software development collaboration ties that emerge between countries through the uncoordinated activities of the hundreds of thousands of individual software developers are a potentially important aspect to consider for understanding international economic interdependence in today's information economy. Furthermore, the structure of OSS collaboration is important for technological innovation, since these mutual cross-border collaborations directly affect the extent of joint-innovations, and their global diffusion potential [1]. As such, gaining a deeper understanding of the principles that govern bottom-up open-source software development across national borders can aid in identifying novel collaboration opportunities for innovation.

2.2 Hierarchy in Global Open-Source Collaboration

A widespread belief about OSS development is that it can remove geographic barriers to enable global collaboration & innovation at an unprecedented rate. However, the reality is that OSS development also clusters geographically [26]. Earlier studies in Linux [5], SourceForge [7, 27], and GitHub [25] consistently indicate that OSS development is dominated by developers from the U.S., the EU, as illustrated in striking visualizations [10]. Even within this concentration of OSS activity and collaborations among Western countries, studies also highlight the extreme hierarchy and asymmetry between the U.S. and other European nations, where software projects led by U.S. developers attracts exceptional levels of international contributions relative to the contributions that U.S. developers tend to make in non-U.S. projects [25]. Although, some recent studies [20, 28] show that geographic concentration in development activity has been gradually decreasing since the 2000s (i.e., participation of non-U.S. and non-EU countries), the projects based in the dominant countries tend to reject code contributions from less developed countries [6, 17, 18].

2.3 Deep-Seated Cultural and Historical Forces

With the unmistakable U.S. dominance in open-source software collaboration, it is difficult not to lose sight of the macro-historical forces underlying international politics, trade, and conflict. For example, the ideological schisms of the Cold War era between the Eastern and Western blocks had a broad and enduring impact on global interactions, but the U.S. dominance in OSS appears to forcefully erase them.

However, political scientist, Samuel Huntington posited that the fault lines in post-Cold War international relations will shift from ideological to deep-seated cultural and religious divides [12]. These divides proved to be important in the formation of distinct blocs and networks, influencing not only political alliances but also scientific, technological, and industrial exchanges. This historical context sets the stage for understanding the paradoxical role of modern communication technologies in global connectivity, as explored in [24]. Challenging the notion that advancements in communication technologies inherently lead to a more connected world, [24] reveals that despite the global reach of digital communication platforms, interactions often reflect and sometimes exacerbate existing cultural divides. The findings suggest a preference for in-group communication, even in digital spaces, indicating that technological tools alone are insufficient for narrowing cultural gaps.

These works collectively paint a picture of global interdependence that is far more complex than the OSS narrative of digital integration and untethered connectivity might suggest. They illustrate how economic, ideological, and technological factors, while having the potential to bring nations closer, also maintain or even widen existing divides. Such a nuanced understanding of global interdependence is crucial for considering the dynamics of international collaboration, particularly in fields like open-source software development. Hence, moving beyond general-purpose communications, we test the effects of broad cultural/religious commonalities on the structure of the specialized, goal-driven collaboration network in GitHub.

3 Methods

3.1 Data

We use the publicly available GitHub Innovation Graph dataset[1] which provides the complete quarterly country-to-country *Git push*[2] volume data from Q1, 2020 to Q2, 2023. GitHub is the most widely used open-source software platform, with more than half of developers reporting its use for both personal and professional projects[3]. Hence, the activities on GitHub can be regarded as representative of the prevailing patterns of open-source collaboration. The approximate locations of developers are inferred from the IP addresses associated with a developer's Git pushes, which are then aggregated to construct country-to-country collaboration ties.[4]

We study the aggregate collaborations across all quarters by creating a weighted, directed graph consisting of countries as nodes and their aggregate Git pushes from one country to another as weighted edges between them.[5] However, this raw graph is heavily

[1] https://innovationgraph.github.com/

[2] The GitHub platform uses the Git version control system. In Git parlance, contributions to a source code repository are packaged as "commits" that get "pushed" to that repository.

[3] https://survey.stackoverflow.co/2022/#overview

[4] https://github.com/github/innovationgraph/blob/main/docs/datasheet.md

[5] The replication data and code are available at: https://github.com/hehao98/github-innovation-graph

skewed with weights between major countries that are an order of magnitude larger than the weights between smaller, developing countries. Given our goal of delineating the multi-scale international collaborations while also filtering out the possible noise observed in the low-weight edges, we first normalize the weight of each edge by the total outgoing weight of its source node, then retain only the edges that are more likely to have structural significance, using the disparity filter for uncovering the network backbone [22].

The disparity filter examines the weights of a node's outgoing ties and retains the edges with weights that are significantly larger, relative to when the weights were to be distributed at random. The filter is customizable with an alpha value at which statistical significance is defined. Choosing a low alpha value results in more aggressive edge filtering. Instead of filtering out any edge with weight below a fixed threshold (i.e., uniform thresholding), the disparity filter adapts to the local structure of the network, preserving the heterogeneity of connections.

3.2 Hierarchical Clustering

To characterize the hierarchical structure of the global OSS collaboration network and identify structural equivalence classes (block models), we employ hierarchical clustering analysis [29]. This approach groups countries based on the similarity of their collaboration patterns rather than their direct connections, revealing the underlying positional structure of the network consistent with world systems theory [23].

We construct structural equivalence by measuring the Euclidean distance between countries' collaboration profiles. For each pair of countries i and j, we compute:

$$d_{ij} = \left\|\mathbf{a}_i - \mathbf{a}_j\right\|_2 = \sqrt{\sum_{k=1}^{n}(A_{ik} - A_{jk})^2} \tag{1}$$

where $\mathbf{a}_i$ denotes the ith row of the adjacency matrix A, representing country i's collaboration profile, and n is the number of countries. Countries with similar distance values have comparable patterns of collaboration with other countries, indicating structural equivalence regardless of whether they directly collaborate with each other [29].

Using the complete linkage method, we perform hierarchical agglomerative clustering on the distance matrix. The optimal number of clusters is determined by maximizing the Calinski-Harabasz score, which evaluates cluster separation and compactness [3]. Formally, the Calinski-Harabasz (CH) index for k clusters and N countries is defined as:

$$\mathrm{CH}(k) = \frac{\mathrm{BCSS}/(k-1)}{\mathrm{WCSS}/(N-k)} \tag{2}$$

where the between-cluster sum of squares (BCSS) and within-cluster sum of squares (WCSS) are computed as:

$$\mathrm{BCSS} = \sum_{i=1}^{k} n_i \left\|\mathbf{c}_i - \bar{\mathbf{a}}\right\|_2^2, \quad \mathrm{WCSS} = \sum_{i=1}^{k}\sum_{\mathbf{a}\in C_i} \left\|\mathbf{a} - \mathbf{c}_i\right\|_2^2 \tag{3}$$

Here, n_i is the number of countries in cluster i, $\mathbf{c}_i$ is the centroid of cluster i, and $\bar{\mathbf{a}}$ is the global centroid of all collaboration profiles. All distances are measured in the same Euclidean space as defined above. A higher CH score indicates more distinct and compact clustering. We evaluated clustering solutions across threshold values from 20 to 150 with unit intervals, selecting the configuration with the highest CH score.

The resulting block model partition identifies structural positions within the global collaboration hierarchy [23]. Countries in the same block share similar roles in the network: they may not collaborate directly with each other, but they maintain comparable patterns of relationships with countries in other blocks. This structural equivalence approach reveals the core-periphery dynamics where the core countries (typically advanced economies) receive disproportionate collaboration from semi-peripheral and peripheral countries, while peripheral countries exhibit sparse internal connectivity despite occupying similar structural positions [23].

This block modeling approach conceptually differs from community detection methods that identify densely connected groups. Instead, it reveals the positional structure of countries based on their equivalent roles in channeling international OSS collaboration flows, providing insight into the hierarchical organization of global technological networks consistent with the theoretical predictions of world systems [23].

3.3 Exponential Random Graph Models

To quantitatively assess the influence of cultural factors on cross-national OSS collaboration patterns while accounting for the complex interdependencies inherent in network data, we employ Exponential Random Graph Models (ERGMs) [19]. Traditional statistical approaches that assume independence among observations are inappropriate for network analysis, as the formation of one collaborative tie may directly influence the probability of forming additional ties within the same network structure [29]. ERGMs address this fundamental limitation by explicitly modeling the conditional dependence structure of network formation processes.

ERGMs belong to the exponential family of probability distributions and model the likelihood of observing a particular network configuration $\mathbf{y}$ as:

$$P_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{y}) = \frac{1}{\kappa(\boldsymbol{\theta})}\exp\left\{\boldsymbol{\theta}^T \mathbf{g}(\mathbf{y})\right\} \tag{4}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$ represents the vector of model parameters, $\mathbf{g}(\mathbf{y}) = (g_1(\mathbf{y}), \ldots, g_p(\mathbf{y}))^T$ denotes the vector of sufficient statistics capturing various network configurations, and $\kappa(\boldsymbol{\theta})$ is the normalizing constant:

$$\kappa(\boldsymbol{\theta}) = \sum_{\mathbf{y}^* \in \mathcal{Y}} \exp\left\{\boldsymbol{\theta}^T \mathbf{g}(\mathbf{y}^*)\right\} \tag{5}$$

where $\mathcal{Y}$ represents the sample space of all possible networks on n nodes.

The conditional log-odds of edge formation between nodes i and j, given the rest of the network $\mathbf{Y}_{ij}^c$, can be expressed as:

$$\mathrm{logit}(P(Y_{ij} = 1|\mathbf{Y}_{ij}^c)) = \boldsymbol{\theta}^T \delta(\mathbf{y})_{ij} \tag{6}$$

where the sum is over all configurations A that contain Y_{ij}, and $\delta_A(\mathbf{y})$ is the change statistic representing the change in the value of the network statistic $g_A(\mathbf{y})$ when y_{ij} changes from 0 to 1.

Network Statistics and Model Terms. Our ERGM specification incorporates several theoretically motivated network statistics to test hypotheses regarding cultural homophily, reciprocity, and structural patterns in international OSS collaboration [11]:

Edges Term: The fundamental density statistic controls for the baseline propensity of collaboration:

$$g_{\text{edges}}(\mathrm{y}) = \sum_{i,j} y_{ij} \tag{7}$$

This statistic counts the total number of collaboration ties in the network.

Nodematch Terms: To test Huntington's cultural homophily hypothesis [12], we employ nodematch statistics that quantify the tendency for countries sharing civilization membership to exhibit higher collaboration rates [14]. For civilization attribute C, the differential nodematch statistic for civilization k is:

$$g_{\text{nodematch},k}(\mathrm{y}) = \sum_{i,j} y_{ij}\mathbf{1}(C_i = C_j = k) \tag{8}$$

where $\mathbf{1}(C_i = C_j = k)$ is an indicator function equal to 1 when both countries i and j belong to civilization k. This specification allows for civilization-specific homophily coefficients, enabling us to test whether different cultural groups exhibit varying degrees of preference for internal collaboration.

For structural equivalence effects, we also test block model homophily using:

$$g_{\text{block model}}(\mathrm{y}) = \sum_{i,j} y_{ij}\mathbf{1}(B_i = B_j) \tag{9}$$

where B_i represents the block model position of country i derived from hierarchical clustering of structural equivalence patterns.

Mutual Terms: For directed networks, reciprocity effects are captured through mutual dyad statistics [29]:

$$g_{\text{mutual}}(\mathrm{y}) = \sum_{i,j} y_{ij}y_{ji} \tag{10}$$

This statistic counts pairs of countries that collaborate with each other in both directions (mutual collaboration).

Baseline Model Specification and Estimation. Our baseline model specification evaluates the structure of the collaboration network while controlling for density and reciprocity effects:

$$\Pr(\mathrm{Y} = \mathrm{y}) = \frac{1}{k(\boldsymbol{\theta})} \exp\left\{\theta_{\text{edges}}L(\mathrm{y}) + \theta_{\text{mutual}}M(\mathrm{y})\right\} \tag{11}$$

where $L(\mathrm{y}) = \sum_{i,j} y_{ij}$ is the number of edges, $M(\mathrm{y}) = \sum_{i,j} y_{ij}y_{ji}$ is the number of mutual dyads.

Parameter estimation is conducted using Monte Carlo Maximum Likelihood Estimation (MCMC-MLE) as implemented in the `ergm` package [9]. The MCMC algorithm approximates the intractable normalizing constant through iterative simulation from the model distribution. We employ a burn-in period of 40,000 iterations with sampling intervals of 10,000 iterations to ensure adequate mixing and convergence. Model adequacy is assessed through goodness-of-fit diagnostics that compare observed network statistics to distributions generated from the fitted model.

3.4 Node2Vec

We test the robustness of both the hierarchical clustering of countries in structurally equivalent positions and the direct collaboration ties based on homophily in a unified node embedding framework, node2vec. Node2vec learns continuous vector representations of nodes by performing biased random walks that can be controlled by two hyperparameters [8]. The return parameter p controls the likelihood of revisiting nodes, while the in-out parameter q controls whether the random walk explores locally or globally. High p and low q biases the random walk toward depth-first exploration, capturing direct collaboration patterns between countries embedded in dense clusters (i.e., homophily). In contrast, low p and high q biases the random walk toward local breadth-first search, capturing countries that play similar structural positions (i.e., structural equivalence) such as hub positions in the collaboration network.

After generating the embeddings with different return parameter p and in-out parameter q combinations from values [0.25, 0.5, 1, 2, 4], we settled with $p = 4, q = 0.25$ for node features learned from homophily and with $p = 0.25, q = 4$ for structural equivalence. We then standardized each embedding using StandardScaler and ran k-means clustering with six clusters using Euclidean distance as the similarity metric. The optimal number of clusters was determined using silhouette analysis [21]. While $k = 3$ and $k = 2$ achieved the highest silhouette scores for the homophily and structural equivalence embedding configurations (0.348 and 0.823, respectively), these provided insufficient granularity for meaningful analysis of the complex global collaboration structure, obscuring important regional groupings and nuanced structural relationships between countries. We selected $k = 6$, which yielded the next-highest scores (homophily: 0.337; structural equivalence: 0.451) while providing adequate resolution to identify distinct collaboration patterns.

3.5 Colonial Dimension Analysis

Building on the node2vec embeddings described above, we investigate whether historical colonial relationships persist in the OSS collaboration network structure by constructing a colonial dimension in embedding space. This approach tests whether countries exhibit systematic proximity to their historical colonizers when positioned along a theoretically motivated vector space derived from colonial relationships.

Specifically, we construct a "colonizer-colonized dimension" in the learned node2vec embedding by computing the average vector from three major historical colonial relationships. Using colonial pairs India-Britain (IN-GB), Vietnam-France (VN-FR), and Mexico-Spain (MX-ES), this colonial dimension is defined as:

$$CD = \frac{(v_{IN} - v_{GB}) + (v_{VN} - v_{FR}) + (v_{MX} - v_{ES})}{3} \tag{12}$$

where v_i represents the node2vec embedding for country i computed using the parameters specified in Section 3.4. This dimension captures the average directional vector from former colonizers toward their colonies, theoretically representing a "decolonized" direction in the embedding space.

We test this dimension by projecting countries of interest along the colonized vector ($v_i + CD$) and measuring cosine similarity to potential colonizers. Test countries and regions include Ukraine, Poland, South Korea, Taiwan, Argentina, Brazil, Canada, Senegal, Turkey, United States, and China. The results are aggregated across three different node2vec parameter combinations (structural equivalence p=0.25, q=4; homophily p=4, q=0.25; balanced p=1, q=1) on the collaboration networks with vs. without the U.S., resulting in a total of six configurations.

4 Results

4.1 Global Hierarchy of OSS Collaboration

The hierarchical clustering analysis reveals four distinct structural positions in the global OSS collaboration network. The dendrogram in Figure 1 demonstrates how countries naturally cluster based on their structural equivalence patterns, with the resulting blocks reflecting real-world geopolitical hierarchies.

The dendrogram shows distinct clustering patterns across all country groups. Among the periphery countries (shown in orange), we observe clear regional groupings where African countries cluster together, Latin American countries form another coherent cluster, and small island nations and developing economies occupy separate branches. The dendrogram structure indicates that despite their geographic dispersion, these countries share similar structural positions in the global OSS network—they predominantly contribute to projects in core countries while receiving minimal collaboration from other periphery nations.

The dendrogram also illustrates the hierarchical relationships among core and semi-periphery countries. The core countries (shown in red) include major Western European nations (Netherlands, Germany, France, United Kingdom), along with economically advanced countries like Canada and Spain. These countries occupy similar structural positions as regional hubs that attract substantial collaboration from their respective spheres of influence. The semi-periphery countries (shown in green) display more heterogeneous clustering patterns, reflecting their intermediate position between core and periphery. Notable clusters include Eastern European countries (Poland, Czechia, Romania), Asian emerging economies (Taiwan, South Korea, Singapore), and other middle-income countries that serve as bridges between the global core and their regional peripheries.

The blue lines in the dendrogram highlight the unique structural position of the United States, which separates early in the clustering process, confirming its exceptional role in the global OSS collaboration network that we analyze in detail below.

The global collaboration pattern exhibits a strict hierarchy among four groups of countries that assume structurally equivalent positions, or "blocks", obtained from a block model analysis. Figure 2 plots the asymmetry in collaborations among these four blocks – core, semi-periphery, periphery blocks and the U.S. as its own block. The cells represent a column block's relative collaboration volume directed to a row block (e.g., the collaboration volume from developers in peripheral countries on U.S. projects is 233.78% of the U.S. developers' collaboration volume on the projects of the peripheral countries). The volume of collaborations directed between clusters are highly asymmetric and transitive – as shown in Figure 2, the core countries consistently receive disproportionately more collaborations from the semi-periphery and the periphery countries while the semi-periphery countries receive out-sized collaborations from the periphery countries. This perfectly transitive structure remains

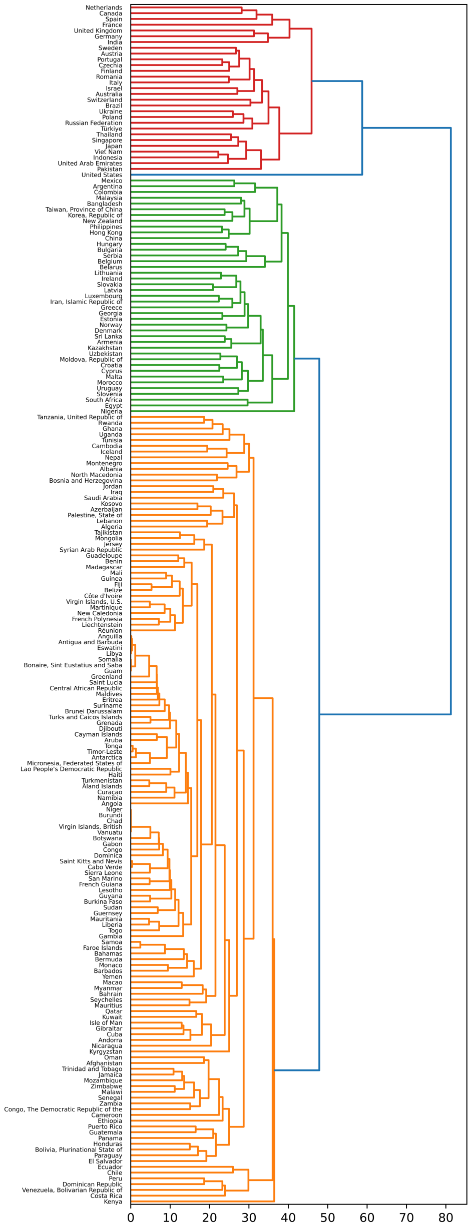


Figure 1: Hierarchical clustering dendrogram of all countries in the global OSS collaboration network. The blue lines highlight the separation of the U.S. from other countries, reflecting its unique structural position. Countries are colored by their structural equivalence blocks: periphery (orange), core (red), semi-periphery (green), and the U.S. as a separate block. The optimal cutting points for the four-block partition were determined by maximizing the Calinski-Harabasz score, revealing the hierarchical organization of the global OSS network.

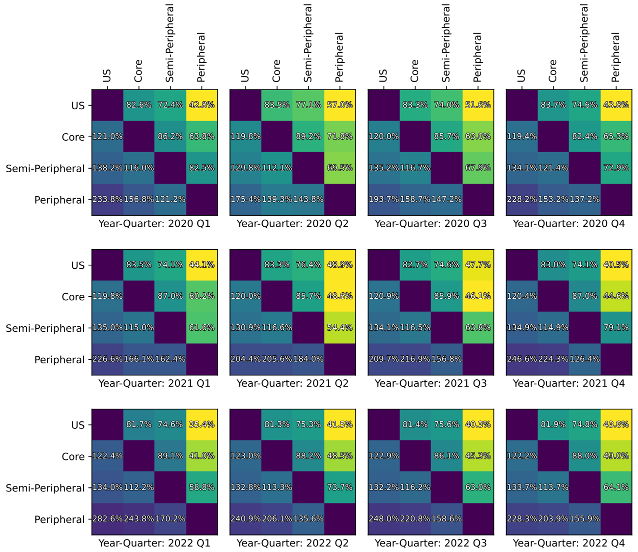


Figure 2: Relative collaboration tie strength between core, semi-periphery, and periphery country positions

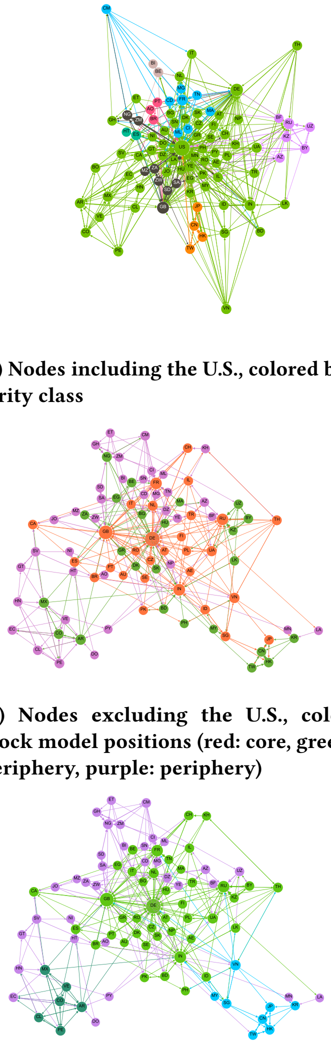
highly stable across quarters with the U.S. at the apex, receiving disproportionate collaborations from the other three blocks.

Indeed, the strong centripetal force of the U.S. is visually apparent in Figure 3a, which displays a subset of countries that appear in Samuel Huntington's civilization classification, based on a force directed layout (Force Atlas 2 in Gephi [2]). Node colors represent countries within densely knit communities of strong collaboration ties, as labeled by the Louvain community detection algorithm. Consistent with previous studies that report the exceptional U.S. prominence in OSS development, it looms large at the center, so much so that over 60% of the countries are labeled into the same community as the U.S., gravitating en masse, thereby blurring the more subtle boundaries.

However, once the U.S. is removed from the graph, the overall hierarchy obtained from the aforementioned block model is more apparent. In Figure 3b with the U.S. removed, the node colors are based on the structurally equivalent blocks of the core (red), semi-periphery (green), and periphery (purple) countries. In the absence of the U.S., the large core countries (i.e., Great Britain, Germany, France, Russia, and India) are spatially dispersed, with their respective semi-periphery countries closely positioned. Aligned with this finding, a qualitatively similar core-periphery structure is apparent from K-means clustered ($K = 6$) countries based on the node2vec embeddings that prioritize structural equivalence ($p = 0.25$, $q = 4$), as shown in Figure 3c. While the embeddings differentiate some of the Latin American and Asian countries as separate clusters respectively, they offer a coarser visual differentiation between the core (green) and periphery (pink).

4.2 Cultural Homophily

We present four ERGMs that incrementally evaluate the influence of structural equivalence and cultural homophily on cross-national



(a) Nodes including the U.S., colored by modularity class

(b) Nodes excluding the U.S., colored by block model positions (red: core, green: semi-periphery, purple: periphery)

(c) K-means clustering (6 clusters) on node2vec embeddings emphasizing structural equivalence (p=0.25, q=4)

Figure 3: Force Atlas layout of the GitHub collaboration graph including the U.S. and colored by modularity class (a), without the US colored by block model positions (b), and k-means clustering applied to node2vec embeddings that emphasize structural equivalence patterns (c). Node size and edge thickness are proportional to weighted nodal degree and edge weight, respectively. While the original k-means clustering identified six clusters, only four are visible in this visualization as the countries in the fifth and sixth clusters were excluded due to weak, non-significant collaboration ties that would have appeared as isolates after applying the disparity filter.

Table 1: ERGM Models 1 & 2: Baseline and Structural Equivalence

Term	Model 1	Model 2
Edges	-3.206^{***}	-3.263^{***}
Mutual	2.294^{***}	2.283^{***}
Nodematch.SameBlock	–	$0.165^{\cdot}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

OSS collaboration. Each model builds on the baseline network structure while testing additional hypotheses, emphasizing cultural affinity and positional equivalence as sources of clustering.

Model 1: Baseline Network Structure. The baseline model includes only the number of edges and mutual collaboration ties. The results in Table 1 show a strong and significant mutuality effect ($\theta = 2.294, p < 0.001$), indicating that dyads with bilateral contributions are substantially more likely to form. However, it is a stretch to interpret mutual as reciprocity in the sense of quid-pro-quo between countries. Since ties are aggregates of individual behaviors, contributors are unlikely to coordinate country-to-country exchanges. Rather, the mutual effect likely reflects broader similarity patterns between countries that are not captured by other covariates.

Model 2: Structural Equivalence (Same Block). Model 2 adds a nodematch term for countries in the same structural equivalence block derived via hierarchical clustering. Table 1 shows that the coefficient is positive but only marginally significant ($\theta = 0.165, p < 0.1$). According to world systems theory [23], countries occupying peripheral positions should exhibit low internal connectivity, with stronger outward links to the core. Hence, the limited magnitude and marginal significance of the same-block effect is consistent with the expected hierarchical structure of global collaboration.

Model 3: Cultural Homophily (Civilizations). Model 3 evaluates the influence of shared cultural affiliation by including nodematch terms for Huntington's civilization categories. Table 2 shows that most civilization terms are strongly significant and positive. The Sinic civilization exhibits the strongest internal collaboration tendency ($\theta = 2.344, p < 0.001$), followed by Orthodox ($\theta = 1.725, p < 0.001$) and Latin American ($\theta = 1.657, p < 0.001$) civilizations. In contrast, Islamic civilization shows no significant homophily effect ($\theta = 0.313$, n.s.), while the African civilization demonstrates a marginally significant effect ($\theta = 0.775, p < 0.01$). These results are visually apparent in Figure 4a where countries in the same civilization are spatially clustered closely together, similar to the unsupervised clustering based on modularity as shown in Figure 4b. Mirroring the weak cultural homophily of African and Islamic nations in Model 3, The African and Islamic countries are spatially scattered in contrast to other tightly clustered non-Western civilizations in Figure 4a (e.g., Latin American, Orthodox).

Model 4: Cultural Homophily and Structural Equivalence Combined. Model 4 incorporates both civilization categories and the positional equivalence of the same block. As shown in Table 2, the civilization coefficients remain robust and significant, while the same block term becomes negative and nonsignificant ($\theta = -0.114$, n.s.), in contrast to Model 2 (Same Block only: $\theta = 0.165, p < 0.1$). This reinforces the interpretation that structural equivalence captures countries with similar positions in the global system, not necessarily countries with stronger bilateral collaborations. Two countries in the same block may or may not have direct ties, because being in the same block simply means that they have similar connection patterns to others (i.e., structural equivalence), not that they should have a higher propensity to form direct ties with each other. According to world systems theory [23], the periphery block should have low internal connectivity, with high interconnections within the core block, fewer within the semiperiphery, and sparse connections within the periphery itself. In short, Model 4 shows that cultural homophily, operationalized through shared civilizational identity, offers more consistent and interpretable insight into direct collaboration ties in global OSS production.

Table 2: ERGM Models 3 & 4: Cultural Homophily

Term	Model 3	Model 4
Edges	-3.518^{***}	-3.490^{***}
Mutual	1.718^{***}	1.730^{***}
Nodematch.SameBlock	–	-0.114
Nodematch.African	0.775^{**}	0.841^{**}
Nodematch.Buddhist	1.902^{***}	1.939^{***}
Nodematch.Islamic	0.313	0.320
Nodematch.LatinAmerican	1.657^{***}	1.690^{***}
Nodematch.Orthodox	1.725^{***}	1.707^{***}
Nodematch.Sinic	2.344^{***}	2.384^{***}
Nodematch.Western	1.604^{***}	1.630^{***}

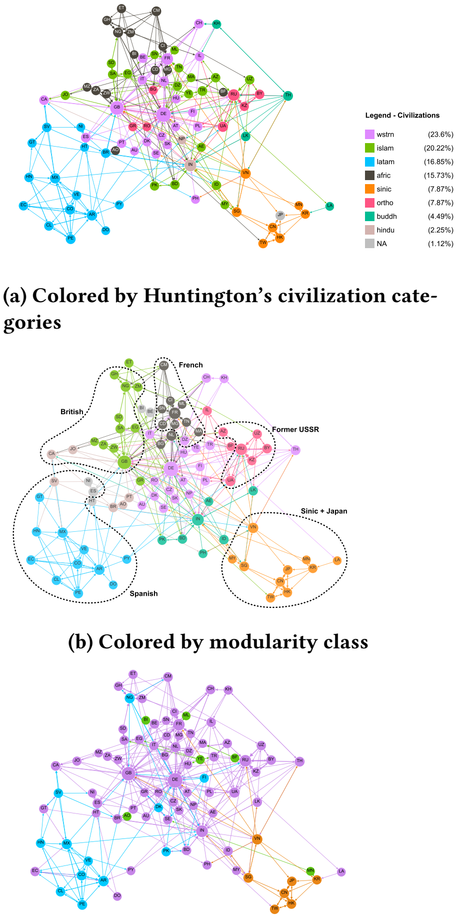
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Finally, the salience of macro cultural similarities for OSS collaboration in Figure 4a is also visually apparent when compared to Figure 4c, which colors the countries according to the clusters ($K = 6$) obtained from the node2vec embedding that emphasizes homophily ($p = 4$, $q = 0.25$). These clusters do not effectively differentiate the sampled countries as the civilization labels, demonstrating the utility of a theory-driven exploration.

4.3 Colonial History and Geopolitics

The unsupervised Louvain community detection, unlike the theory-driven groupings of countries based on structural equivalence (Figures 3b, 3c) or homophily (Figures 4a, 4c), additionally hints at historical factors that continue to structure OSS collaborations. Of particular note, the modules shown in Figure 4b, obtained from community detection, coincides with groups of countries that share common colonial histories, such as the former British colonies located in the Northwest, neighbored by the former French colonies in the North, and by the former Spanish colonies to the Southwest. The former Soviet bloc countries are also grouped as one community.

We analyzed the relationships between the colonizers (France, UK, Spain) and the cold-war super powers (Russia) on the one hand and their subordinated countries on the other by projecting the node vectors on the colonial dimension, as described in Section 3.5. The results provide direct evidence of macro-historical legacies



(a) Colored by Huntington's civilization categories

(b) Colored by modularity class

(c) K-means clustering (6 clusters) on node2vec embeddings emphasizing homophily (p=4, q=0.25)

Figure 4: Force Atlas layout of the GitHub collaboration graph without the U.S., colored by Samuel Huntington's civilization labels (a), modularity class (b), and k-means clustering applied to node2vec embeddings that emphasize homophily patterns (c). Node size and edge thickness are proportional to weighted nodal degree and edge weight, respectively. While the original k-means clustering identified six clusters, only four are visible in this visualization as the countries in the fifth and sixth clusters were excluded due to weak, non-significant collaboration ties that would have appeared as isolates after applying the disparity filter.

embedded within the collaboration network structure. Japan consistently emerges as the closest to South Korea and Taiwan (4/6 configurations), accurately reflecting Japanese colonial history. Spain matches Argentina in 3/6 configurations and Portugal matches Brazil in 2/6 configurations, suggesting the persistence of Iberian colonial legacies in Latin American OSS collaboration. Beyond formal colonialism, Russia frequently appears closest to Ukraine and Poland, reflecting Soviet sphere of influence, analogous to the above-mentioned traditional colonial relationships.

However, contemporary superpowers appear to defy historical projection. Both the United States and China consistently exhibit highest similarity to themselves (>0.99 cosine similarity), indicating these nations maintain such dominant structural positions that even colonial dimension adjustments cannot overcome their network centrality.

Taken together, the apparent organization of collaboration around common colonial history (and cold war hierarchies in the Soviet block) may partly explain why countries from the Islamic and African cultures are less visually clustered within their own culture, but scattered around their respective former colonizers. The economic dependencies that date back to the colonial era continue to influence today's knowledge economy that is supported in part by OSS development.

5 Discussion

Extending previous studies that identify the geographic locus of OSS development, this study dissects the network structure of the cross-national collaboration on Github. As reported in the literature, we observe a strictly hierarchical collaboration structure with the U.S. as the strongest gravitational force at the epicenter of the global collaboration network, followed by a core group of technologically advanced countries heavily contributing to U.S.-based projects. The U.S. and this core group, in turn, hosts projects to which the semi-periphery and periphery countries heavily contribute. This transitive structure in terms of core, semi-periphery, and periphery countries is reminiscent of the hierarchical structure of the international trade network that social network analysts have repeatedly discovered from input-output tables since the 1970s [23].

Furthermore, these international collaborations also appear to reflect long-term cultural affinities and shared historical experiences between countries. By triangulating the results from network visualizations, block models, exponential random graph models, and node embeddings, we demonstrated the robust influence of deep-seated cultural and religious similarities (i.e., civilizations) as well as long-lasting geopolitical power structures from the past (i.e., colonial power-dependencies and cold-war super powers). These results are not likely to simply reflect shared language uses between countries, given that only 12.7% of Github developers used non-English comments in their code commits [16].

Nevertheless, the comparison across different analytical approaches reveals both the robustness and underlying complexity of collaboration patterns. Both hierarchical block modeling and k-means clustering based on node2vec biased toward structural equivalence identify the pluralistic core, consisting of DE, GB, FR, RU, and IN and their close collaborating nations, surrounded by the diverse peripheral and semiperipheral countries broadly positioned on the outskirts, along with two additional geographically clustered blocks of Latin American and Asian countries, respectively. In contrast, while the modularity-based community detection accurately distinguishes colonial spheres, the k-means clustering based on node2vec embeddings learned with the homophily biased parameters do not detect these fine distinctions as clearly as the community detection approach. Discrepancies not withstanding, these results from triangulating different methods offer a nuanced illustration of how homophily and structural equivalence are manifested at multiple scales within both historical networks and emerging regional collaborations.

5.1 Discussion

While the findings of this study are informative, they are to be evaluated with the following data limitation. The reliance on IP-based activity statistics for constructing the weighted collaboration edges could introduce potential biases, particularly in countries with prevalent VPN usage, like China. This could lead to distorted representations of actual OSS activity. While we anticipate these distortions are limited in scope, their impact on the study's findings should not be underestimated. Future research might explore alternative data collection methods to mitigate this limitation. Despite this limitation, our findings offer several key insights with broad implications:

5.1.1 Influence on Commercial Software Development Practices. Our results indicate that historical and geopolitical factors have a significant impact on OSS contributions. They suggest new ways for imagining and evaluating global outsourcing and collaboration strategies in commercial software development, which could lead to tangible development of agile and adaptable collaboration models for global software teamwork.

5.1.2 Overcoming Barriers in OSS Collaboration. As we have demonstrated, identifying and addressing both inter-community and intra-community barriers in OSS collaboration can be challenging. Future initiatives might focus on developing enhanced recommender systems for OSS platforms, incorporating cultural tags to foster more effective cross-cultural collaborations that are conducive to innovation [15]. Additionally, there is an opportunity to develop features within platforms like GitHub to streamline collaboration within culturally similar groups, potentially evolving these platforms into more effective multi-hub network structures.

5.1.3 Broader Research Horizons. Future research might also explore the socio-political dimensions of OSS collaboration, examining how global political shifts and policy changes influence the dynamics of these networks. Understanding these broader impacts could provide valuable insights into the evolving nature of global software development and collaboration.

6 Conclusion

All in all, the implications of this study are far-reaching, suggesting the need for resilient OSS community frameworks, adaptable commercial software development strategies, and tools to strike the right balance between the efficiency gains of hierarchical integration at the global scale while also lowering both inter and intra-community collaboration barriers.

Acknowledgement

This work is supported in part by a Google faculty research gift.

Generative AI Disclosure

No generative AI tools were used in any stage of this research, including data collection, analysis, code development, and manuscript preparation.

References

[1] Tufool Alnuaimi, Jasjit Singh, and Gerard George. 2012. Not with my own: long-term effects of cross-country collaboration on subsidiary innovation in emerging economies versus advanced economies. *Journal of Economic Geography* 12, 5 (2012), 943–968. http://www.jstor.org/stable/26158628

[2] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

[3] T. Caliński and J Harabasz and. 1974. A dendrite method for cluster analysis. *Communications in Statistics* 3, 1 (1974), 1–27. doi:10.1080/03610927408827101 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/03610927408827101

[4] Kevin Crowston, Kangning Wei, James Howison, and Andrea Wiggins. 2012. Free/Libre open-source software development: What we know and what we do not know. *ACM Comput. Surv.* 44, 2 (2012), 7:1–7:35. doi:10.1145/2089125.2089127

[5] Bert J. Dempsey, Debra Weiss, Paul Jones, and Jane Greenberg. 2002. Who is an open source software developer? *Commun. ACM* 45, 2 (2002), 67–72. doi:10.1145/503124.503125

[6] Leonardo B. Furtado, Bruno Cartaxo, Christoph Treude, and Gustavo Pinto. 2021. How Successful Are Open Source Contributions From Countries With Different Levels of Human Development? *IEEE Softw.* 38, 2 (2021), 58–63. doi:10.1109/MS.2020.3044020

[7] Jesús M. González-Barahona, Gregorio Robles, Roberto Andradas-Izquierdo, and Rishab Aiyer Ghosh. 2008. Geographic origin of libre software developers. *Inf. Econ. Policy* 20, 4 (2008), 356–363. doi:10.1016/J.INFOECOPOL.2008.07.001

[8] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. *CoRR* abs/1607.00653 (2016). arXiv:1607.00653 http://arxiv.org/abs/1607.00653

[9] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2008. statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *Journal of Statistical Software* 24, 1 (2008), 1–11. doi:10.18637/jss.v024.i01

[10] Brandon Heller, Eli Marschner, Evan Rosenfeld, and Jeffrey Heer. 2011. Visualizing collaboration and influence in the open-source software community. In *Proceedings of the 8th International Working Conference on Mining Software Repositories, MSR 2011 (Co-located with ICSE), Waikiki, Honolulu, HI, USA, May 21-28, 2011, Proceedings*, Arie van Deursen, Tao Xie, and Thomas Zimmermann (Eds.). ACM, 223–226. doi:10.1145/1985441.1985476

[11] 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. doi:10.18637/jss.v024.i03

[12] Samuel P. Huntington. 1993. The Clash of Civilizations? *Foreign Affairs* 72, 3 (1993), 22–49. http://www.jstor.org/stable/20045621

[13] Matthew O Jackson and Stephen Nei. 2015. Networks of military alliances, wars, and international trade. *Proceedings of the National Academy of Sciences* 112, 50 (2015), 15277–15284.

[14] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, Volume 27, 2001 (2001), 415–444. doi:10.1146/annurev.soc.27.1.415

[15] Scott Page. 2007. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition)* (01 2007).

[16] Chris Piech and Sami Abu-El-Haija. 2020. Human Languages in Source Code: Auto-Translation for Localized Instruction. In *Proceedings of the Seventh ACM Conference on Learning @ Scale* (Virtual Event, USA) *(L@S '20)*. Association for Computing Machinery, New York, NY, USA, 167–174. doi:10.1145/3386527.3405916

[17] Ayushi Rastogi, Nachiappan Nagappan, and Georgios Gousios. 2016. *Geographical bias in GitHub: Perceptions and reality.* Technical Report. https://repository.iiitd.edu.in/jspui/bitstream/handle/123456789/388/IIITD-TR-2016-001.pdf?sequence=3

[18] Ayushi Rastogi, Nachiappan Nagappan, Georgios Gousios, and André van der Hoek. 2018. Relationship between geographical location and evaluation of developer contributions in GitHub. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2018, Oulu, Finland, October 11-12, 2018*, Markku Oivo, Daniel Méndez Fernández, and Audris Mockus (Eds.). ACM, 22:1–22:8. doi:10.1145/3239235.3240504

[19] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. 2007. An introduction to exponential random graph (p*) models for social networks. *Social networks* 29, 2 (2007), 173–191.

[20] Davide Rossi and Stefano Zacchiroli. 2022. Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years. In *19th IEEE/ACM International Conference on Mining Software Repositories, MSR 2022, Pittsburgh, PA, USA, May 23-24, 2022.* ACM, 80–85. doi:10.1145/3524842.3528471

[21] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. doi:10.1016/0377-0427(87)90125-7

[22] M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. 2009. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences* 106, 16 (April 2009), 6483–6488. doi:10.1073/pnas.0808904106

[23] David Snyder and Edward L. Kick. 1979. Structural Position in the World System and Economic Growth, 1955-1970: A Multiple-Network Analysis of Transnational Interactions. *Amer. J. Sociology* 84, 5 (1979), 1096–1126. doi:10.1086/226902 arXiv:https://doi.org/10.1086/226902

[24] Bogdan State, Patrick Park, Ingmar Weber, and Michael Macy. 2015. The Mesh of Civilizations in the Global Network of Digital Communication. *PLOS ONE* 10, 5 (05 2015), 1–9. doi:10.1371/journal.pone.0122543

[25] Yuri Takhteyev and Andrew Hilts. 2010. Investigating the geography of open source software through GitHub. *Manuscript Submitted for Publication* (2010). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=9ba29373aac78aa592f3cbf932fbd2d14d6fcb53

[26] Babu Veeresh Thummadi and Srikanth Paruchuri. 2022. Presence of Location-Based Agglomeration Effects in Open Source Communities: An Empirical Test on GitHub. *Academy of Management Discoveries* 8, 2 (2022), 274–297. https://doi.org/10.5465/amd.2019.0255

[27] Sebastian von Engelhardt, Andreas Freytag, and Christoph Schulz. 2013. On the Geographic Allocation of Open Source Software Activities. *Int. J. Innov. Digit. Econ.* 4, 2 (2013), 25–39. doi:10.4018/JIDE.2013040103

[28] Johannes Wachs, Mariusz Nitecki, William Schueller, and Axel Polleres. 2022. The geography of open source software: Evidence from GitHub. *Technological Forecasting and Social Change* 176 (2022), 121478. https://doi.org/10.1016/j.techfore.2022.121478

[29] Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.