UNIVERSITY OF CHICAGO

# Open Particles in Social Medium: Open Collaboration in GitHub

By

Shiyang Lai

July 2023

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: James Evans
Preceptor: Shilin Jia

2023

TABLE OF CONTENTS

TABLE OF CONTENTS.....II
ACKNOWLEDGEMENTS..... III
ABSTRACT..... IV
INTRODUCTION.....V
LITERATURE REVIEW.....XII
DATA.....XX
INSTRUMENTS OF ANALOGICAL ANALYSIS.....XXIV
SOCIAL ATOMS' SUPERPOSITION STATE.....XXXVI
DISCUSSION.....XLII
REFERENCES.....XLVIII
APPENDIX A..... LII
APPENDIX B.....XLVIII

# ACKNOWLEDGEMENTS

First and foremost, my profound appreciation goes out to my advisor, Prof. James Evans. Your academic guidance, constant patience, and relentless encouragement have been pivotal in this journey. Your wisdom lit the way in the course of my research, for which I am deeply grateful.

Next, I would like to express my heartfelt appreciation to my MA preceptor, Dr. Shilin Jia, along with Professors John Padgett, Andrew Abbott, Luc Anselin, Anton Strezhnev, Chris Graziul, and Ningzi Li. Their invaluable feedback, enlightening suggestions, and profound academic wisdom have greatly enhanced the development of this thesis. Their generous contribution of knowledge has nurtured a dynamic academic milieu, fostering significant growth on my part as a scholar.

Lastly, my sincere gratitude goes to my family and friends, for their steadfast belief in my capabilities and their constant support and understanding during the challenging stages of this journey.

I am eternally grateful to everyone for their invaluable contribution to my journey.

# ABSTRACT

This thesis proposes an innovative approach for the comprehensive and structured modeling of online open collaboration. By conceptualizing individuals, teams, and collaborative environments as social atoms, molecules, and mediums respectively, it employs concepts and tools from physics and chemistry to illustrate human interactions. This research introduces an analogical framework that retains the fundamental principles of select traditional natural science instruments while also integrating more probabilistic and dynamic elements to address the disparities between the physical and social systems. It applies and adapts two tools and one concept from physics and chemistry - namely the collaborative periodic table (inspired by the periodic table of elements), the probabilistic collaboration equation (based on the reaction equation), and the social superposition state (derived from quantum physics). These tools and concepts are employed to study open collaboration on GitHub, leveraging a dataset sourced from the GitHub open-source community that includes collaborative data from hundreds of thousands of users. The findings of this study not only provide deeper insights on the behavior patterns of open collaboration participants, but also illuminate a bottom-up strategy for future sociophysical pursuits.

# INTRODUCTION

Over the last two decades, we have witnessed the flourishing of a vast list of "open" projects that are developed through unbounded organizations. *Wikipedia*, a well-known case of such projects, which is produced mostly by anonymous end users, has come to match the quality of encyclopedia texts (e.g., *Encyclopædia Britannica*) that have been refined by credentialed experts over years (Giles, 2005). Open source software is an even more exhilarating example. Exemplified by *Linux* and *Android* operating systems, they have become commonplace, running on most smart devices and creating trillions of dollars in economic value (Levine & Prietula, 2014). Constitutionally, the success of all "open" projects is anchored in a new sociotechnical system of harnessing the *wisdom of the crowd*, which we now formally term *open collaboration*.

Yet, today's open collaboration remains in its nascent stage, with participants continuing explore its new applications and optimal practices across distinct domains of organizational activity (Splitter et al., 2023). The context of open collaboration is experiencing significant transformations as well. Technologically, the maturation of remote collaboration software and the advancements in AI are eradicating the myriad of physical and cultural barriers that impede traditional collaboration (Bernstein & Turban, 2018; Botsman & Rogers, 2010; Ferraro & O'Mahony, 2012; O'Mahony & Bechky, 2008; Yang et al., 2022). Socially, an ever-growing number of professionals and learners from diverse fields are flocking to open collaboration communities, giving rise to cooperative networks that are more diverse and loosely structured than conventional organizational frameworks. Amidst this ever-evolving landscape, the growth of open collaboration has become a perplexing

phenomenon, encompassing not only its expanding magnitude but also its intricate form which is likely to be much more complicated than traditional organizations due to less structural constraints and the weak-tie nature of such collaborations. This provides a meaningful yet challenging opportunity for us to explore and model the underlying dynamics that govern the interactions and behaviors of individuals and groups within open collaboration communities. With this context in mind, my study aims to present a new analytical framework based on a socio-physical perspective, allowing for a more systematic and in-depth examination and modeling of open collaboration.

The framework adopts a physicochemical analogy to conceptualize indispensable components in human collaboration. Specifically, it maps individuals, teams, and the collaborative environments onto social atoms, molecules, and mediums, respectively. This approach draws inspiration, in part, from the social-chemical comparison presented by Ethat et al. as well as a long list of works on sociophysics (Abergel et al., 2017; Arnopoulos, 2005; Bernstein & Turban, 2018; Galam, 2012; Parongama Sen & Bikas K. Chakrabarti, 2014; Pentland, 2015). It is worth noting that the employed analogy here may be particularly appropriate for open collaboration. In contrast to traditional organizational collaboration settings, where employees are mostly *passively selected* to collaborate, open collaboration participants have much more freedom to *voluntarily choose* their work engagements. This unique characteristic renders collaboration predominantly a self-selection process, bearing greater resemblance to physical particle interactions, and thus making the analogical approach feasible in this unique context (Baldwin & von Hippel, 2011). Besides, there are two other benefits of drawing parallels between human collaboration and physical particle interaction. First, the causal processes inherent in the micro-level physical particles'

interaction system exhibit a striking resemblance to those in social collaboration. For instance, individuals transitioning from spatially constrained workspaces to remote work settings, which fosters extensive large-scale open collaboration, is analogous to the shift of atoms from a solid to a liquid medium, which stimulates the generation of new chemical reactions. Second, this approach facilitates *model transfer*, a scientific practice that involves repurposing a model initially applied to a specific target system within a particular scientific domain to represent a new target system in a different domain (Tan, 2023). Model transfer has been widely recognized as a prominent approach for scientific knowledge production (Humphreys, 2019; Lin, 2022). In this specific investigation, the notions of the periodic table and reaction equation will be showcased as instrumental tools for constructing models of open collaboration and deepening our comprehension thereof.

Nonetheless, two potential caveats of a straight socio-physical analogy should always be awared. First, the physical mechanisms are characterized by determinism, as atoms possess well-defined categorizations with clear boundaries (e.g., metals and noble gases), and the outcomes of chemical reactions, given specific initial conditions, are highly predictable. Conversely, social collaboration encompasses a greater degree of randomness and uncertainty. Assigning individuals to clear-cut groups may be inappropriate, as they can exhibit diverse characteristics in various contexts, and no stable governing laws can justify deterministic human collaboration patterns. Furthermore, the dynamics of human collaboration are considerably more significant. The development of individuals over time plays a critical role in comprehending collaboration, as these changes can occur over relatively short periods. In contrast, alterations in the physical properties of atoms are largely inconsequential, given that the radioactivate decay takes over hundreds of years.

In light of the above considerations, this study proposes a generalized analytical framework that adheres to the fundamental principles of certain classical instruments from natural sciences, while simultaneously incorporating more probabilistic and dynamic elements into its formulation to resolve the two raised caveats. Two specific instruments from physics and chemistry are transferred and socialized in this study, that is the *collaboration periodic table* (transferred from elements periodic table) and *the probabilistic collaboration equation* (transferred from reaction equation). Both instruments are designed to be built upon observational data. To construct the collaboration periodic table, an exploratory factor analysis (EFA) is conducted first to enable a two dimensional representation for social atoms[1], and then, the dynamic Mixed Membership Stochastic Blockmodel (dynMMSBM) introduced by Olivella and associates is employed to further classify them, like how natural scientists categorize elements into various types such as metals and noble gases, but in an probabilistic form instead (Olivella et al., 2022). The collaboration periodic table provides deeper insights to open collaboration patterns by elucidating the heterogeneous nature of the constituent collaborators.

The *probabilistic collaboration equation* is developed in three stages. First, network categorization. In this stage, open collaboration projects with similar team structures and composed of similar types of members are grouped together. These groups represent different types of social molecules that emerge as outputs of the probabilistic collaboration equation. Second, predictability checking, which focuses on assessing the accuracy of predicting the group-ship of a social molecule based on the atomic categories of its constituents. This is

[1] To maintain consistency with the established analogy, the subsequent section of this paper employs the terms "social atoms" for individuals, "social molecules" for teams, and "social mediums" for collaboration platforms.

achieved by utilizing the established collaboration periodic table. Third, fitting probabilistic collaboration equation. The collaboration equation was defined as a multinomial logistic model. Regression analysis was then conducted to estimate the composition of "typical" social atoms for each type of social molecule. Examining these probabilistic collaboration equations holds great significance as it provides deeper insights into the distinct collaboration behaviors exhibited by social atoms, thus enhancing the meaningfulness of the artifact periodic table. Note that the integration of uncertainty within social atoms and social molecules also facilitates the modeling of their dynamism, achieved merely by permitting alterations in their probabilistic representations over time. At this point, the foundations for the analytical approach become solidified. Figure 1 provides a graphical diagram of the elaborated construction[2].

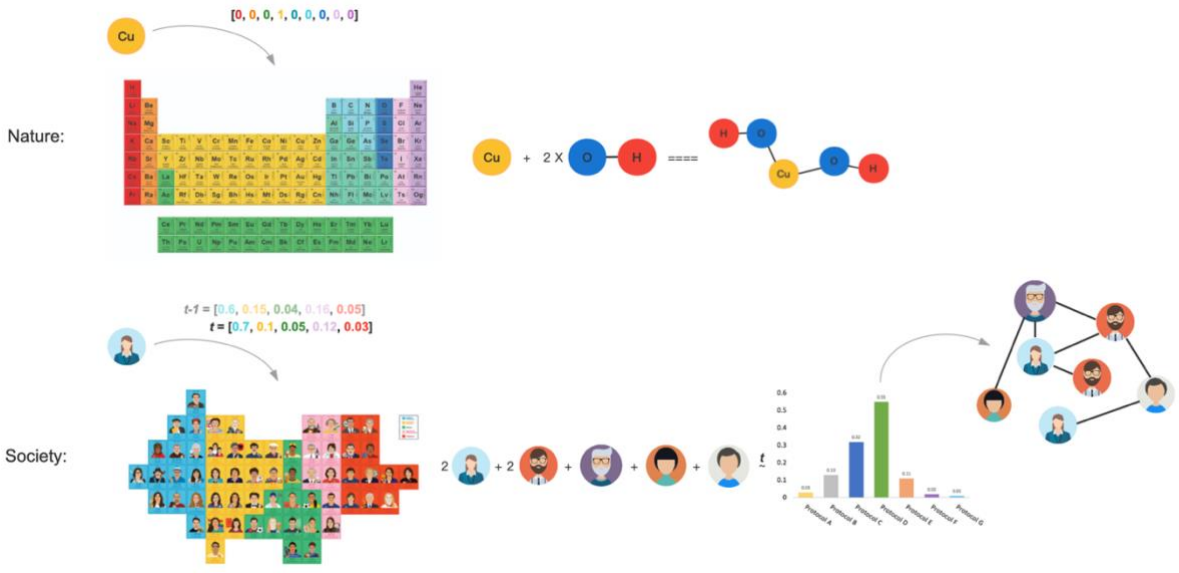


Figure 1: A generalized analogical framework for analyzing open collaboration

[2] The social periodic table depicted in Figure 1 is purely illustrative of the framework, and does not carry any substantive meaning. The image source is https://www.amazon.com/Carson-Dellosa-Set-Periodic-Featuring-Homeschool/dp/B08QYYQ3S1?th=1.

Moreover, in an effort to enhance the prediction of open collaboration participants' behavior and elucidate its inherent uncertainty, this study incorporates the notion of "social superposition states," an idea inspired by quantum physics. The foundational argument is that a social atom does not serve as the most elemental unit in the socio-physical system. Rather, a single social atom encapsulates a spectrum of distinct social states. For instance, during online collaboration, both offline and online social states coexist within a social atom, each contributing its unique traits and behaviors, jointly shaping the atom's overall behavior. This implies that the behaviors of participants in online open collaboration are likely shaped by a combination of their virtual and physical surroundings. To provide empirical support for this claim, a spatial econometric analysis was undertaken in this study to examine whether individuals' online work patterns are co-influenced by their digital and geographical environments.

The application and scrutiny of this analytical methodology for the reexamination of open collaboration are founded upon an extensive, self-compiled dataset of GitHub open-source repositories. This dataset comprises metadata for 168 thousand repositories, accompanied by the particulars of their 416 thousands contributors and 7 million pertinent historical activity records. the subsequent sections of this paper are organized as follows: Section 2 provides overviews on open collaboration and sociophysics. Section 3 outlines the data collection and preprocessing steps. Section 4 details the construction of the collaboration periodic table and the probabilistic collaboration equations. In Section 5, particular attention is devoted to the empirical examination of the superposition principles manifested by social particles, which demonstrate how the framework extends to the subatomic level. Finally, Section 6 presents a synopsis of the findings and concludes with some closing observations.

# LITERATURE REVIEW

## *Open Collaboration: Its Birth, Conceptualization, and Social Implications*

There is much discussion about the changing nature of work in response to rapid technological advances. Conversational media such as social networking websites, question-and-answer sites, social coding platforms, and remote meeting software have built a digital double of the real-world physical workspace, and then the long-lasting COVID-19 pandemic further forces people to really adapt remote work mode in recent three years (McLean, 2020; Yang et al., 2022). Living in such an era, managers and organizational scholars have been increasingly framed organizational boundaries as barriers that ought to be spanned, permeated, and blurred to enable better collective knowledge production (Bernstein & Turban, 2018). Thus, exercises of "expanding" and "removing" collaboration boundaries have been carried out across many disciplines. For software development, most public projects today grant their end users the rights to study, modify, and redistribute their publicly accessible source code on social coding platforms, thereby turning the development process to commons-based peer production (Levine & Prietula, 2014). Meanwhile, many businesses use dedicated crowdsourcing sites like to find solutions to niche tasks like graphic design, proofreading, and software testing. In the academic sphere, open science and citizen science are booming. Researchers now engage volunteers to code large data sets, participates in online experiments, create children's books, translate ancient texts, and much more (Eklund et al., 2019; Franzoni & Sauermann, 2014). Deboundarying efforts from all these sectors come together, forming the landscape of open collaboration in this very age.

The formal definition of open collaboration is firstly conceived by Riehle et al. in

2009 to refer to collaboration principles in open-source software projects. In their definition, open collaboration conforms with three principles, namely egalitarianism, meritocracy, and self-organization (Riehle et al., 2009). Egalitarianism suggests the equal accessibility for anyone who want to contribute, even they are complete strangers; meritocracy means that contributions are judged transparently on the basis of merits; self-organization indicates the absence of predefined process outside, or say, open collaboration should be autonomously operated. Riehle et al.'s definition highly summarizes the features of collaboration in open-source software development but is also narrowly restricted to this particular context. Later, Forte and Lampe devised a much more flexible definition of open collaboration, which sticks with Riehle's first principle, egalitarianism, and substitutes the second and third with one consequential feature, which claims that open collaboration fosters the emergence of persistent yet malleable social structures. This new definition made a longer list of collaborative activities fall into the family of open collaboration and, more prominently, highlighted its social dimensions (Forte & Lampe, 2013). Subsequent definitions of open collaboration by other scholars have largely been consistent with Forte and Lampe's conception, although they may differ in terminology and emphasis due to diverse interests (Mergel, 2015; Scaliza et al., 2022).

As Forte and Lampe highlighted, open collaboration cultivates persistent yet adaptable social structures. Unlike traditional organizations, which rely on formal social contracts to regulate social structures and establish hierarchical relationships with clear power dynamics and lines of authority, open collaboration online allows for voluntary and low-cost establishment of relationships. This results in the majority of social connections being weak ties (Yang et al., 2022). Granovetter argues that weak ties are essential because they act as

*bridges* between different social groups, facilitating the exchange of information and resources between them (Granovetter, 1973). Thus, open collaboration enables individuals to collaborate with others outside of their immediate niches, leading to increased social capital for participants and promoting a more diverse and inclusive social environment.

Early to the 20th century, people have already realized that the ethos of open collaboration aligns with the values and principles of modern science (David, 1998). This alignment stems from two facts. First, the norm of openness is conducive to the collegiate reputation-based reward system, where credit is given to those who establish priority in their claims. Second, it facilitates strategic choices that reduce the duplication of research efforts and enlarge the domain of information complementariness. However, Baldwin and Hippel contend that science remained confined to a traditional "producers" model until the advent of digital-mediated collaboration platforms pivoting scientific innovation to a "users" model, indicating that true openness was not achieved until then (Baldwin & von Hippel, 2011). They claim that online open collaboration leads to a reorientation of the problem field in science, from a "discipline-centered mode" to a "question-centered mode". This transformation catalyzes conversation across paradigms as it becomes costless for people from different niches today to reunion and debate for the same question of interests. During this process, traditional paradigms are cross-validated and new paradigms are established over the traditional ones, and, ultimately, fostering the next wave of scientific evolution (Kuhn & Hacking, 2012). On the other hand, open collaboration induces a transformation of the power-knowledge nexus (Foucault, 1966). In the milieu of open collaboration, the real-life social status and racial identity of individuals are obscured, leading to a more egalitarian milieu for generating knowledge. Yet, novel power dynamics arise surrounding online

reputation and media algorithms, which subsequently impact the production and diffusion of knowledge.

Despite its potential advantages, open collaboration has been associated with various negative impacts as well. Unlike traditional collaboration, open collaboration operates with indistinct membership and porous boundaries, where anyone can consume without contributing. As a result, free riding and unequal contribution become common outcomes (Levine & Prietula, 2014). But counterintuitively, by simulation, researchers show that open collaboration can still prosper even in environments dominated by non-contributors (Baldwin & von Hippel, 2011; Levine & Prietula, 2014). Another concern regarding open collaboration is that it may lead to conflict and disagreement, as it lacks clear lines of authority and decision-making structures. Consequently, the diversity of contributors may frustrate groups, as conflicts of opinion and vision may be challenging to resolve in 'headless' organizations (Du Chatenier et al., 2009; Wang et al., 2021). Open collaboration can also amplify the spread of misinformation, particularly in the absence of formal quality control measures. This can lead to confusion and a lack of trust among participants, as well as a decline in the group's overall effectiveness (Avieson, 2022).

As previously mentioned, open collaboration carries extensive ramifications. However, a discernible shortage of empirical research examining the practices and consequences of open collaboration persists. Factors contributing to this gap include the limited availability of pertinent data sources and the lack of suitable research methodologies for quantitatively modeling the complex social dynamics inherent in open collaboration. In light of this, the present study endeavors to address this research void by proposing a generalized approach to modeling open collaboration that is applicable in diverse contexts.

The proposed methodology is anticipated to yield novel empirical insights, thereby substantially enriching the existing body of literature on open collaboration.

### *Sociophysics: Gaining Inspiration from Nature Sciences*

The analogy approach of this study, generally, lies under the umbrella term “Social Physics”, which was first introduced by the Belgain statistician Adolphe Quetelet in 19th century (Quetelet, 1835). Subsequent to Quetelet's introduction of the term, a growing number of statistical physicists have been inspired to apply physical methodologies and analytical paradigms to quantitatively assess and study social phenomena (Cho, 2009). This has led to many innovative developments, some of which have been very revolutionary. For instance, researchers have drawn an analogy between the formation of public opinion through individual social interactions and the alignment of magnetic fields in neighboring atoms in a crystal. This analogy inspired them to employ variants of the Ising model as a means to depict this unique form of social interaction (Galam, 2013; Ishii & Okano, 2021; Schelling, 1971; Stauffer, 2002). Another influential research area, largely influenced by sociophysics, focuses on complex networks - a favored method for portraying intricate social systems. In these networks, social agents are depicted as nodes, and their interactions as links. The systemic attributes are then accounted for by the structure of interactions, or in other words, by the topology of the network (Abergel et al., 2017; Jusup et al., 2022; Parongama Sen & Bikas K. Chakrabarti, 2014; Schweitzer, 2018).

Prior to delving into the primary tenets of sociophysics, it is imperative to explicate the general tactical procedure it employs. The most prevalent modus operandi within sociophysics aligns with Galam's phenomenon-centric method as elucidated in his 2012

publication. This methodology commences with the selection of an intriguing phenomenon or social practice, with subsequent pinpoint identification of the paradoxical features affiliated with it. Following this, the next steps involve the isolation of the paramount parameters of the identified paradox and selecting appropriate physical models in an effort to elucidate and effectively decode the intrinsic paradoxical elements (Galam, 2012). There have been plenty instances exhibiting the application of this particular style of sociophysics in the last ten years (Abergel et al., 2017; Parongama Sen & Bikas K. Chakrabarti, 2014), yet this method has not been without its detractors. Critiques have emerged from varied audiences spanning both the social and natural sciences disciplines, indicating a breadth of interdisciplinary concern.

The critiques on the mainstream approach of sociophysics can be generally classified into two major groups. The first concerns the questionable relevance of physical models to social dynamics. The application of physical tools to the exploration of social science questions often relies on sociophysical analogies such as “social atoms”, which may not be wholly suitable (Cho, 2009; Jensen, 2019). As articulated by Castellano et al., the models employed by physicists begin with isolated “simple entities” possessing stable traits, and the goal is to discern the macroscopic regularities emerging from the interactions amongst these entities. However, humans, unlike atoms, are continuously reshaped by our interactions and experiences. The absence of a stable core that fundamentally defines us and governs our actions raises doubts about the reliability of models that predict collective behavior based on the assumption of static utility functions across individuals or over time (Castellano et al., 2009). For instance, in Schelling's model, individuals are characterized solely by their color and their utility function, attributes that remain unchanged throughout the model's duration

(Schelling, 1971).

The second major category of detractors contends that contemporary sociophysics overly simplifies the intricate weave of social interactions, reducing them to rigid mathematical models. These models often fall short in accounting for the numerous influential factors and the fluid, unpredictable nature of social settings (Parongama Sen & Bikas K. Chakrabarti, 2014). Physics, in its elegance, conforms to fixed and deterministic laws that can be beautifully encapsulated in mathematical models. For instance, the gravitational force between two objects can be accurately forecasted using the succinct formula $F = \frac{Gm_1m_2}{r^2}$, which incorporates just four factors. However, the application of such an approach to social systems poses challenges. Consider, for example, the attempt to predict voting behavior. It might seem logical to craft a mathematical model based on quantifiable variables like age, gender, income level, and education. However, these models are often inadequate, as voting decisions are also shaped by less quantifiable elements such as a person's values, emotions, candidate perceptions, social circle influence, and even the information consumed in the lead-up to the election (Dennison, 2019). Therefore, the adoption of physical models to unravel social science problems may not yield much additional insight. This is primarily due to the omission of substantial complexity intrinsic to social systems, a complexity that these models often fail to encapsulate but carries great significance within the realm of social sciences.

Besides, there are physicists claiming that today’s sociophysics suffers from a lack of empirical substantiation. They argue that the theoretical premises established in this field often lack the backing of data or fail to be verified through experimental research (Vazquez, 2022). Some other critics also suggest that physicists transitioning into the social sciences

may not entirely comprehend the intricacies and sophistication of social theory, resulting in overly simplistic analyses of social phenomena (Schweitzer, 2018).

Responding to the aforementioned critiques, this research presents an alternate object-oriented framework for social physics. To narrow the disconnect between physical models and social phenomena, and rectify the oversimplification typically seen in physical models, the framework starts by forming a direct correspondence between fundamental units of analysis within the specific social scenario and comparable entities within the physical system. Recognizing their similarities, the next crucial step involves identifying the disparities between each analogous pair. Subsequently, the associated tools and concepts are modified and generalized to enhance their compatibility when applied within the social system. The framework is data-driven, meaning the adjusted social physics concepts are ultimately validated and specified using real-world data. In this way, it offers a unique solution to the existing criticisms of sociophysics.

# DATA

## *Data Collection*

The strategies for obtaining the essential information on repositories, users, and events, which are necessary for investigating GitHub's open collaboration, are briefly expounded here at first. It should be noted that all metadata is gathered through the GitHub REST APIs, a web-based tool that provides access to GitHub's historical public archive. The data collection process begins by retrieving information on repositories established between January 2018 and October 2022, categorized by month and star rating[3]. For instance, to illustrate the approach, I start with January 2018 and identify the public repositories created in that month with the highest number of stars. To lessen the burden on server requests and guarantee comprehensive representation across the entire spectrum of star ratings, data acquisition is capped at 1,000 repositories for each 10-star increment. Ultimately, repositories that created at the first month of 2018 and spanned the entire range of star ratings are collected. I repeat the above steps for each month, resulting in a dataset of GitHub repositories containing a total of 168,407 entries.

Afterwards, data on repository-specific events are collected. The GitHub repository dataset includes URLs that enable requests for complete historical events records for corresponding repositories. Based on this, a repository-event dataset is established, containing detailed information on 7 million historical events. Then, I list all the actors' IDs

[3] Repositories with star ratings are often popular, well-maintained, or otherwise noteworthy. The number of stars can give an at-a-glance estimation of how popular or significant a repository is within the GitHub community.

appeared in the repository-event dataset, a total of 416 thousands GitHub users, and further request their basic information to create the GitHub user dataset. With the user ID, the user-specific historical activity information then can also be obtained. This information is stored in a separate dataset that complemented the existing user dataset. In total, the complement user-event dataset recorded 36 million events.

***Data preprocessing***

To render the extensive and unstructured metadata more conducive to analysis, a thorough cleanup and reorganization of the gathered GitHub repositories, user, and event datasets are undertaken. During the cleanup process, the repository dataset is refined by eliminating repositories that are either forks of others or have fewer than three collaborators; the user dataset excludes individuals who have been active on GitHub for a duration of less than two weeks or have never pushed code; the repo-event and user-event datasets are amended by removing event records linked to deleted repositories and users. Concerning the reorganization, given that the focus of this study lies in collaborative behaviors rather than all activities performed by GitHub users (e.g., creating and deleting repositories), a distinct collaboration link table is generated by retaining only interactive activities such as *pull requests*[4], *pull request reviews*, *pull request review comments*, and *issue comments* in the repository-event dataset.

***Variable Operationalization***

[4] A pull request in GitHub is a procedure to propose changes to a project's codebase. It is interactive in nature because it initiates a collaborative review process where multiple contributors can inspect, discuss, and refine the proposed changes.

Twelve variables are constructed to capture the exhibited characteristics of GitHub users. The definitions of these variables and their basic descriptive statistics are reported in Table 1.

Table 1: Definition and statistical description of user-level variables

| **Continuous Variable** | **Definition** | **Mean** | **Std** | **Min** | **Max** |
|---|---|---|---|---|---|
| *Experience* | Active years on GitHub | 5.859 | 3.800 | 0.000 | 15.129 |
| *Activity* | Average number of acts on GitHub per day | 2.202 | 7.420 | 0.000 | 300.000 |
| *Workday_Work_Intensity* | Average number of push requests per business day | 2.854 | 6.936 | 0.000 | 169.000 |
| *Weekend_Work_Intensity* | Average number of push requests per weekend day | 1.712 | 4.014 | 0.000 | 150.000 |
| *Collaborativity* | Average number of collaborative acts per day | 1.340 | 6.077 | 0.000 | 144.000 |
| *Productivity* | Average number of created repositories per active year | 6.561 | 43.712 | 0.000 | 82.031 |
| *Skill_Diversity* | Natural log of the number of programming languages appeared in their owned repositories | 2.581 | 0.767 | 0.693 | 5.867 |
| *Engagement* | Number of followings | 18.255 | 192.844 | 0.000 | 38720.000 |
| *Reputation* | Number of followers | 59.257 | 602.481 | 0.000 | 90176.000 |
| **Categorical Variable** | **Definition** | **Number of Categories** | | | |
| *Country* | Country of residence | N=178 | | | |
| *Freelancer* | Whether the person is a freelancer | N=2 | | | |
| *Main_Language* | The mostly used programming language | N=309 | | | |
| *Institution* | The belonged institution type (e.g., government and university) | N=7 | | | |

*Note: For variables such as Activity, Workday_Work_Intensity, Weekend_Work_Intensity, Collaborativity, measurements were taken separately for the years 2022 and 2023.*

Eight variables on GitHub open-source repositories are operationalized. The definitions of these variables and their basic descriptive statistics are shown in Table 2.

Table 2: Definition and statistical description of project-level variables

| **Continuous Variable** | **Definition** | **Mean** | **Std** | **Min** | **Max** |
|---|---|---|---|---|---|
| *Size* | The size of the entire repository in KB | 41.930 | 690.543 | 9.000 | 1.026e5 |
| *Popularity* | The number of stars received | 397.275 | 1786.91 | 0.000 | 2.643e5 |
| *Usefulness* | The number of times of being forked | 81.919 | 509.173 | 0.000 | 74398.000 |
| *Activity* | The number of open issues | 14.504 | 105.041 | 0.000 | 19547.000 |

| **Categorical Variable** | **Definition** | **Number of Categories** |
|---|---|---|
| *Main_Language* | The mostly used programming language | N=318 |
| *Domain* | The belonged domain (i.e., data science, software development, education, and finance) | N=4 |
| *License* | The adopted software license | N=38 |
| *Created_Year* | The year of being created | N=5 |

*Note: Domain variable is extracted by performing the Latent Dirichlet Allocation (LDA) model on the self-reported keywords of repositories. The number of domains is determined by the elbow method. See Table A1 in appendix A for the details of the topic modeling.*

### *Collaboration Periodic Table*

This section unveils the initial analogical instrument of the framework — the collaboration periodic table. Drawing parallels to the periodic table of chemical elements, the collaboration periodic table maps distinct types of social atoms to their shared interaction patterns. Its formulation hinges on empirical data, and in this study, I employed a large GitHub dataset to create a specific collaboration periodic table dedicated to participants in GitHub's open-source software development. It is critical to recognize that the table is fundamentally an artifact that can be modified according to its construction methods. In physical and chemical research, the periodic table is also artifact and its significance stems from its ability to condense rich empirical knowledge on the complex physical particle dynamics into a two-dimensional reference table that is both easy to understand and adequately informative. This table enables natural scientists to forecast properties of diverse elements, even unobserved ones, given their known positions on the table. In a similar vein, the creation of the collaboration periodic table isn't focused on unveiling any causal processes intrinsic to human collaboration. The primary drive is to develop a handy reference tool capable of simplifying the even more sophisticated social system, while simultaneously encapsulating its fundamental heterogeneity, commonality, and predictability. The collaboration periodic table allows for analysis of individuals and their interactions at the atomic category and atomic group levels. The empirical knowledge garnered from these analyses further enhance the table, granting it predictive and interpretive capabilities akin to those of the periodic table in natural sciences.

Following an exploratory factor analysis[5], two latent factors for GitHub users are identified and named as the Activity Factor and the Programming Expertise Factor based on their loadings. The Activity Factor and Programming Expertise Factor are rescaled to range from 0 to 1 and evenly partitioned into 15 and 10 intervals. This process results in a $10 \times 15$ grid of atomic cells, where users situated in the same cell are classified as the same type of social atom. In the periodic table of chemical elements, atoms are further categorized into distinct groups, such as metals and noble gases. Ingeniously designed, atoms within the same group consistently cluster together in the periodic table. Drawing an analogy to this, a social atom categorization process is employed to identify social atoms exhibiting similar properties and to investigate whether these broader atom groups also exhibit clustering tendencies within the collaboration periodic table. To do so, the dynMMSBM regression is employed to reveal the latent group structures of GitHub users[6]. The dynMMSBM is a statistical model used in network analysis to capture dynamic community structure and tie formation in temporal networks. The independent variables of dynMMSBM generally involve node features and edge features of the temporal network, as well as the time or stage of the network. The dependent variable of dynMMSBM is typically the structure of the network itself, which can be represented by the adjacency matrix of the network or some other representation of the ties between nodes in the network. Unlike traditional clustering techniques, the

[5] The EFA includes all individual-level variables except *Country* and *Institution*. The number of factors is set to be two based on Kaiser's criterion. The Promax rotation method is adopted. The Activity Factor's three highest factor loadings include *Activity* (1.054), *Collaborativity* (0.663), and *Work_Intensity* (0.549). The Programming Expertise Factor's top three factor loadings consist of *Skill_Diversity* (0.741), *Main_Language_C* (0.302), and *Experience* (0.297).

[6] Due to heavy computation costs, the results reported here is based on a model fitted with 10% of the entire data set. Nevertheless, this experiment is repeated 10 times to ensure the consistency of the results. The optimal latent group number is determined to be 6 by comparing the average held-out AUC scores in a cross-validation experiment.

dynMMSBM considers both social interactions and personal attributes, providing a probabilistic portrayal of social atom classifications grounded in an array of discerned latent categories. The regression results are presented in Table 3.

Table 3: The group-specific dynMMSBM regression results

| **Variable** | **Group 1** | **Group 2** | **Group 3** | **Group 4** | **Group 5** | **Group 6** |
|---|---|---|---|---|---|---|
| *Experience* | -6.158e-4 | 0.038 | -0.019 | -0.002 | 0.024 | 1.623e-3 |
| *Activity* | 4.668e-4 | 0.014 | 0.010 | 0.001 | 0.023 | -1.096e-3 |
| *Workday_Work_Intensity* | 0.019 | 0.102 | -0.011 | -0.002 | 7.859e-3 | -7.439e-3 |
| *Weekend_Work_Intensity* | 0.014 | 0.102 | -0.008 | 0.001 | -2.483e-3 | 0.016 |
| *Skill_Diversity* | -5.543e-3 | -0.011 | -0.012 | -0.007 | -6.675e-3 | 0.013 |
| *Engagement* | 4.870e-5 | 6.507e-5 | -3.163e-4 | 8.316e-4 | -2.723e-4 | 4.001e-4 |
| *Reputation* | -1.579e-4 | -3.573e-4 | -8.106e-5 | -1.689e-4 | 4.246e-5 | 1.290e-5 |
| *Freelancer* | -0.077 | -0.299 | 0.404 | -0.013 | -0.081 | -0.177 |
| Main_Language | Yes | Yes | Yes | Yes | Yes | Yes |
| Institution | Yes | Yes | Yes | Yes | Yes | Yes |
| Portion | 16.6% | 15.0% | 17.8% | 18.0% | 13.8% | 18.9% |

*Note. The regression does not incorporate productivity as a factor due to its potential to cause multicollinearity. The report solely focuses on monadic predictors to ascertain patterns for the formation of collaboration. It's important to note that the standard errors associated with the coefficients are not included in the table. This exclusion is attributed to the constraints pertaining to computing resources.*

The results of the regression analysis shed light on a multifaceted relationship between *Experience* and collaboration tendencies within distinct groups. Specifically, it was found that for groups 1, 3, and 4, there is a negative association between *Experience* and their inclination to collaborate. This implies that as individuals in these groups accrue more experience, their tendency to collaborate appears to decrease. In contrast, the opposite trend is evident for individuals in groups 2, 5, and 6, who display an increased propensity to collaborate with a rise in their experience levels. In terms of *Activity*, an intriguing finding is that only individuals within group 6 exhibit a negative coefficient. This suggests that for this group alone, an increased level of activity corresponds to a decrease in collaboration tendencies. When examining working habits, diverse patterns emerge across the groups. Groups 1 and 2 tend to display an amplified collaborative drive with an increase in work input, while group 3 displays a contrary tendency, with collaboration decreasing with

additional work input. Interestingly, groups 4 and 6 show a decrease in collaboration when working more during regular business days, whereas group 5's collaboration decreases when work is performed during the weekend, marking a distinct trait for this group. A noteworthy observation from the *Skill_Diversity* coefficients across the six groups is that with the exception of group 6, all other groups tend to display an increase in collaboration when they possess a narrower range of skills. This finding suggests a potential relationship between skill diversity and the propensity to collaborate, where a lower level of skill diversity might encourage greater collaboration among most individuals. Furthermore, the results of the analysis elucidate the impact of platform engagement and personal reputation on collaboration behaviors. Specifically, individuals in groups 1, 2, and 6 appear to collaborate more as their engagement with the platform increases, while the opposite trend is observed for individuals in groups 3 and 5. Similarly, a higher reputation seems to enhance collaboration for individuals in groups 5 and 6, whereas the inverse relationship holds true for the remaining groups. Lastly, the freelance status seems to have a positive impact only on group 3, who tend to collaborate more when they are not formally employed. For individuals in the other groups, a lack of formal employment appears to decrease their propensity to collaborate. In sum, the regression analysis reveals an intricate and diverse array of relationships between multiple factors and their influences on collaboration tendencies across different groups.

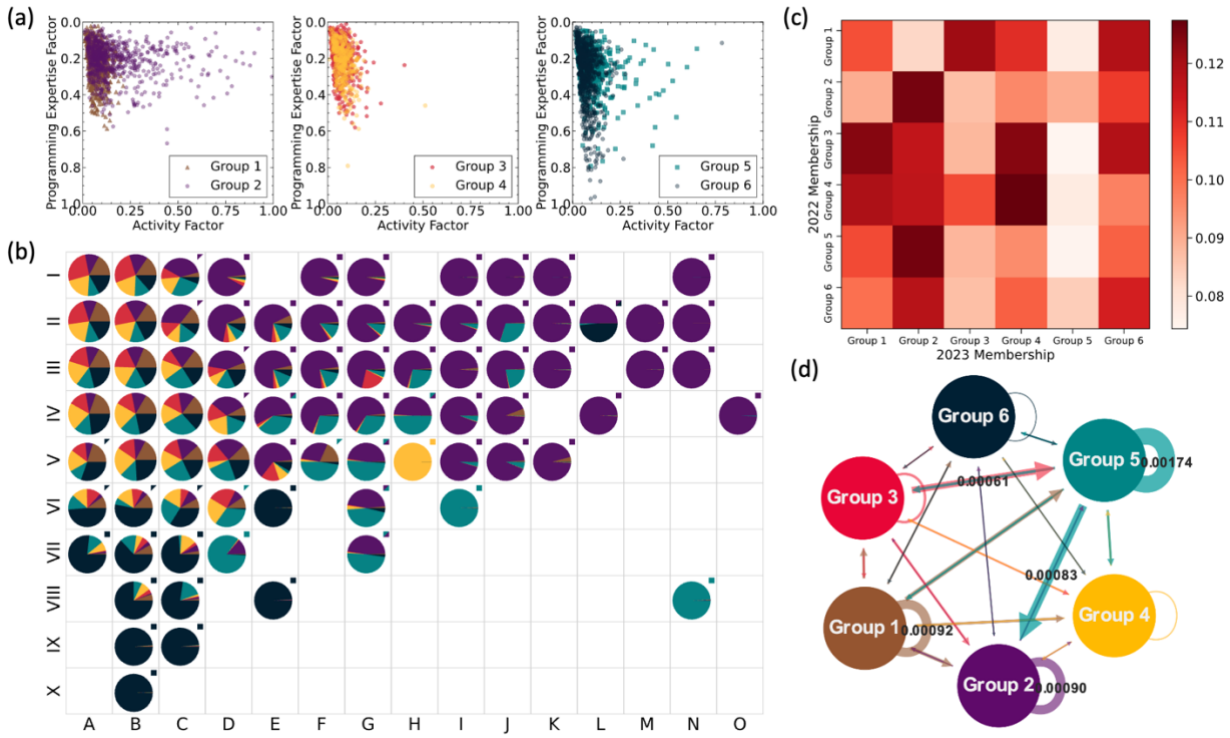


Figure 2**:** The construction of ideal collaboration periodic table

Figure 2 (a) presents a comparison of the distributions of users who are confidently (with at least 80% certainty) assigned to latent groups 1 and 2, 3 and 4, as well as 5 and 6 within the two-dimensional space generated by EFA. As shown in the plots, group 1, 3, and 4 all clustered in the same area of the space, however, the distributions of atoms in group 2, 5, and 6 show distinguishable patterns: GitHub users with high activity and collaborativity levels but low programming expertise tend to belong to group 2; users with low activity and collaborativity levels but high programming expertise are more likely to be in group 6; and users with both high activity and collaborativity levels and programming expertise are more likely to be in group 5.

In a more interpretable collaboration periodic table form, Figure 2 (b) illustrates the proportion of each group of social atoms residing in the same atomic cells using pie charts. If a specific group constitutes over 70% of a cell, the upper right corner of the cell is marked

with a rectangle colored to correspond with that group. If two groups jointly account for more than 70% of the cell, with each group contributing no less than 30%, the upper right corner of the cell is marked with a rectangle filled with both groups' corresponding colors. Similar to the chemical periodic table, atomic cells that are dominated by social atoms in the same latent groups also appear to be within the same regions of the collaboration periodic table. This observation suggests that, despite reducing users' high-dimensional representations to the two extracted factors, their characteristics' heterogeneity is still effectively captured. However, in contrast to the chemical periodic table, the groups in the GitHub collaboration periodic table do not exhibit clear-cut boundaries, a reasonable outcome given the probabilistic nature of social systems. In the case of social atoms situated in the upper left portion of the table, it is difficult to determine their group membership with certainty, as they are more likely to be novices and their behavioral records on the platform are not substantial enough to support a confident classification. The empty cells in the table are places with no observations (i.e., unobserved social atoms).

Figure 2 (c) displays the probability of social atoms transitioning between groups from 2022 to 2023. Intriguingly, the membership of social atoms exhibits robust cross-group dynamics. The colors of the diagonal elements for groups 1, 3, 5, and 6 are not the most intense in their respective rows, signifying that, rather than remaining in the same group in 2023, social atoms belonging to these four groups are more likely to transition to other groups the following year. Even for groups 2 and 4, the probability of staying in the same groups for the next year is only around 13%. This finding underscores the significance of social atoms' dynamism, which is not present in the natural world's atomic counterparts.

Figure 2 (d) displays the blockmodel matrix generated by the dynMMSBM in a network representation, illustrating the likelihood of collaboration occurring within and across groups. Thicker edges between nodes signify a higher probability of forming collaborative relationships. The top five collaboration edges are labeled with the estimated collaboration probability. Three of the five are self-looped edges (i.e., group 1, 2, and 5), which can be attributed to homophily, while the other two indicate a strong tendency for cross-group collaboration. Specifically, social atoms from group 3 demonstrate a pronounced preference for collaborating with social atoms from group 5, and social atoms from group 5, in turn, exhibit a higher likelihood of collaborating with social atoms in group 2.

***Probabilistic Collaboration Equation***

A generic probabilistic collaboration equation is defined to symbolize common dynamics of open collaboration. Referring to the form of chemical reaction equation, it is defined as:

$$WX \overset{C}{\rightarrow} Y$$

*X* refers to a particular group of social particles (reactants). The matrix *W* is composed of "stoisociometric" coefficients, analogous to the stoichiometric coefficients in chemistry, which indicate the number of social particles needed for the resulting outcomes. *Y* is the discrete probabilistic distributions of the outcome social particles. *C* represents the conditions, which encompasses numerous other external factors that can influence the dynamics. This overarching form can be employed to portray all types of collaborative dynamics within an open collaboration system, including team formation (*combination*

*reaction*), team dissolution (*decomposition reaction*), and member transitions between teams (*displacement reaction*).

In this study, I only discuss team formation in the case of GitHub open collaboration. In this special case, X should be a set of social atoms and Y is their combined social molecule. Unlike in the physical world, where molecules are finite in type, teams have an infinite range of possible structures and compositions. This makes the identification of teams with identical collaboration patterns improbable, even when utilizing the simplified social molecule model. To reduce complexity, social molecules on GitHub are further classified into five prototypes based on their internal topological similarities, employing both the NetSimile graph similarity measure and the Louvain Community Detection algorithm (Berlingerio et al., 2012; Blondel et al., 2008). Figure 3 (a) illustrates the categorization. In the network visualization, nodes represent repositories, with their sizes determined by the number of stars received. Their colors correspond to the assigned prototypes. The strength of ties is determined by the similarity between the topologies of the team structures. The five social molecule prototypes are designated as *Complex*, *Parallel*, *Structural*, *Star*, and *Linear* based on their topological characteristics, and they will be considered as the outcomes of the probabilistic collaboration equation. To better show the categorized social molecule prototypes, I randomly selected one repository from each prototype and displayed their collaboration networks.

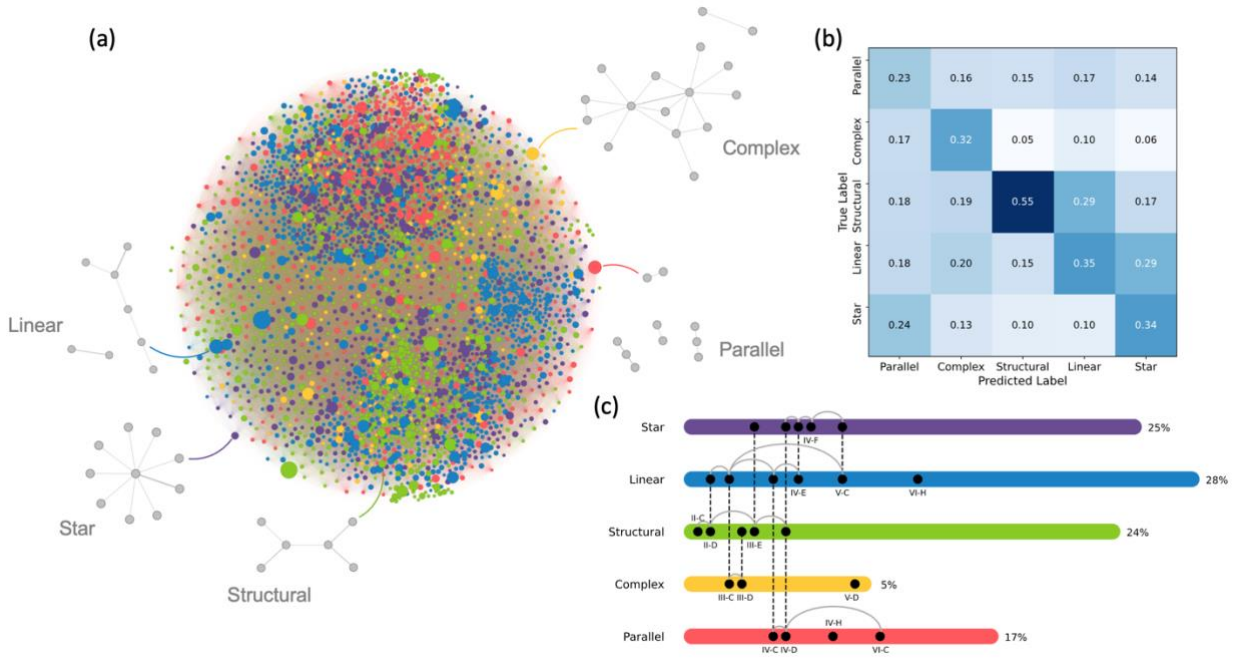


Figure 3: The preliminary analysis for probabilistic collaboration equation

In order to assess the feasibility of formulating a probabilistic collaboration equation for open collaboration pattern formation within the realm of open-source software development on GitHub, it is imperative to first investigate if the knowledge of members' atomic categories genuinely contributes to enhanced predictive capabilities for projects' collaboration prototypes. Figure 3 (b) presents the confusion matrix for the prediction outcomes derived from a simple multi-layer perceptron classifier[7]. Evidently, the prediction accuracy for each prototype surpasses that of a random classifier (i.e., 20%), indicating that the incorporation of individuals' atomic categories indeed provides valuable information for predicting collaboration patterns.

[7] To ensure unbiased prediction performance interpretation, the dataset is first balanced using the SMOTE (Synthetic Minority Oversampling Technique) algorithm. The multilayer perceptron model is then trained for 100 epochs and fine-tuned using a four-fold grid search cross-validation method. The training set counts 80% of the entire dataset, and the presented confusion matrix is normalized by column.

To gain granular insights, I also conducted an association rule mining analysis to mine out what particular social atoms or social atoms' combinations are driven the pattern of open collaboration[8]. The results are visualized in Figure 3 (c). The percentage displayed on the right side of the bars are the proportion of social molecules belonging to the corresponding prototype. Social atoms involved in the top three detected association rules with each prototype as consequent are visualized on their bars. Coexisting social atoms across prototypes are connected with dashed vertical lines. The solid grey curves linking social atoms indicating that the presence of such co-appearance also increase the chance of forming corresponding social molecule prototypes.

The analysis above demonstrates that the atomic categories of social atoms, particularly those displayed in Figure 3 (c), are associated with the structure of the social molecules they form. However, unlike in the natural world, this association is not deterministic in nature. To formalize the probabilistic collaboration equation for team formation on GitHub, I rewrite (1) to the form of a multinomial logistic model.

$$\rho C + WX = Y$$

*X* is a binary column vector $[x_1, x_2, \dots, x_n]^T$, with $x_i$ denoting the presence of social atom *i*, and *n* signifying indicating the total count of distinct social atom types observable. *W* is an $m \times n$ matrix, in which m represents the total number of potential social molecule prototypes. Each $w_{ij}$ element can only take positive values, signifying the anticipated quantity of social atom $i$ within a social molecule that belongs to prototype $j$. *C* is also a column vector $[c_1, c_2, \dots, c_k]^T$, with $c_i$ designating external factor $i$ related to GitHub open collaboration

[8] The algorithm employed for association rule mining is Apriori. Rules are ranked based on the lift metric. A detailed description table can be found in Appendix A, Table A2.

formation, such as the domain (e.g., software engineering, data science) and time. $\rho$ is an $m \times k$ coefficient matrix for *C*. *Y* can be expressed as $[y_1, y_2, \ldots, y_m]^T$, with $y_i$ representing the probability of the open collaboration pattern adhering to prototype $i$. $\rho$ and $W$ can be estimated based on the known *C*, *X*, and $Y^*$ (i.e., the observed outcome prototypes). Figure 4 displays the estimated $\rho$ and *W* matrices. In both matrices, coefficients with p-values below 0.01 are denoted in purple (for negative coefficients) and green (for positive coefficients). A deeper color signifies a larger absolute value of the coefficient. For *W*, red lines are used to divide the matrix by rows of the collaboration periodic table. Social atoms within the same block share similar programming expertise; however, their relative positions in the block represent their activity levels, with those positioned further to the right being more active.

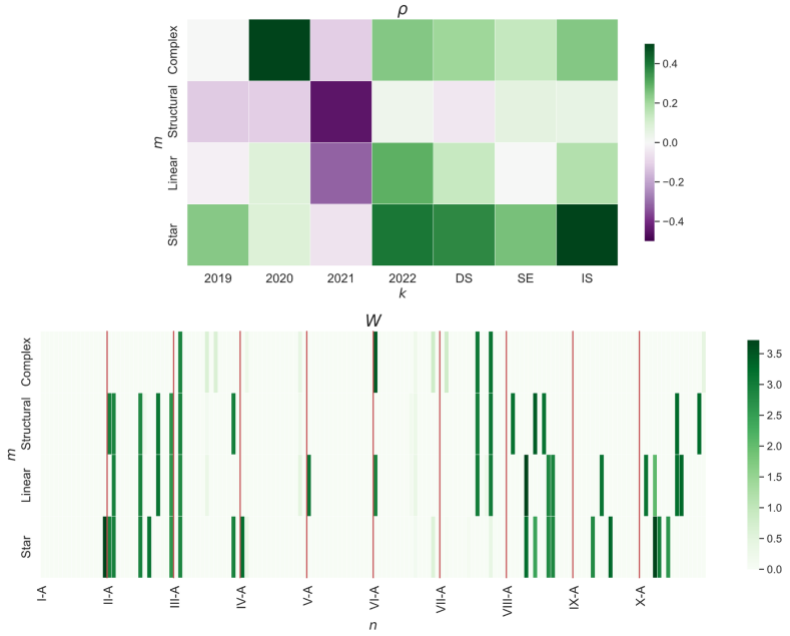


Figure 4: Estimated coefficient matrixes for $\rho$ and $W$

As a noteworthy observation, I delve into some intriguing findings derived from the regression estimations of the probabilistic collaboration equations. The significant "stoisociometric" coefficients suggest an ideal social atom composition for each social

molecule prototype. There are some similarities across prototypes, though, each prototype still retains unique features that set it apart from the others. For *Complex* social molecules, the programming skill expertise of their expected social atoms is less polarized compared to the other three types of social molecules. When a group of social atoms possesses skills that are relatively similar, it becomes less likely for them to establish a clear power hierarchy during interaction due to their comparable capacities. This provides an explanation to the development of *Complex* social molecules. For *Structural* social molecules, it seems that not only their collaboration structures but also their social atom composition is highly structuralized. There are primarily three groups of social atoms expected within a *Structural* social molecule, each group exhibiting clear distinctions in skill expertise. The first group of social atoms is mainly situated between rows II and III, the second group between rows VII and IX, and the third group in row X. For *Linear* social molecules, the skill expertise distribution of the expected social atoms in these types of social molecules is more continuous. This observation is anticipated because, in teams working serially, collaborators with varying levels of programming expertise are needed to facilitate smoother bi-directional communication. Finally, for *Star* social molecules, the skill expertise of their social atoms exhibits a more bi-polarized distribution. In a collaborating group of social atoms with a clear distinction between low and high programming expertise, those with higher capacities are likely to become the central figure of the group, consequently creating a star-like collaboration typology.

## SOCIAL ATOMS' SUPERPOSITION STATE

This section's discussion aims at expanding the analogical framework to subatomic level by introducing an experimental concept, *social superposition state*. The concept of superposition state originates from quantum physics and pertains to the ability of quantum particles such as atoms to exist in multiple simultaneous states or configurations. This notion is exemplified by the renowned thought experiment, Schrödinger's Cat. In a similar vein, social atoms can be likened to Schrödinger's Cats, as each embodies a blend of various *social states*. For instance, the child-state when interacting with parents, the student-state within educational settings, the professional-state within the context of occupations, the friend-state when socializing with peers, etc. Each of these social states manifests distinct characteristics and behaviors. When social atoms are isolated from all social mediums, they can be analogically considered as in a superposition state, a combination of all their diverse social states. Once reintegrated into a specific social context, their uncertain superposition states collapse into more definite states.

However, unlike physical atoms which can only exist in a unique medium at a given time, social atoms can simultaneously inhabit multiple mediums. As a result, unlike physical particles, social atoms do not collapse into a single definite state. Instead, they transition into another superposition state with reduced uncertainty due to the interconnectedness of social mediums. Accordingly, social atoms' superposition states suggest that individuals' presented features are the co-product of the interactions of the multiple layers of mediums they locate. Related empirical evidence has been achieved in recent years (Scellato et al., 2021). In the context of this research, GitHub users, while engaging in online collaboration, concurrently

maintain a physical presence in specific geographical locales. Therefore, in this particular scenario, their online-state and offline-state still interweave, suggesting that their behaviors are likely influenced by both their online and physical environments. To empirically show this, a spatial econometric analysis was conducted to investigate how the surrounding environments of GitHub open collaboration participants in both the GitHub digital medium and physical medium affects their work patterns.

This analysis focused on GitHub users who (1) are residents of the United States, China, Germany, or India, having provided their addresses up to the second administrative area level, (2) have demonstrated activity on GitHub for a period exceeding two weeks, and (3) have submitted at least one push request. To ensure the quality of the data, I also excluded outlier samples[9]. The final selection yields 9,187 observations in China, 5,237 observations in the United States, 6,177 observations in Germany, and 4,210 observations in India. Table A3 provides a descriptive statistics for the variables in the four subsample sets. To normalize the user-reported addresses, I utilized the GPT-3.5 large language model provided by OpenAI to automatically format addresses into the "state/province-county/district" structure. I acquired geographical shapefiles for the four countries from gadm.org. By associating these shapefiles with the four subsample sets, I aggregated the sample sets as a cross-sectional dataset for the subsequent real analysis.

Two spatial regression model specifications were considered. First, taking spatial heterogeneity into account, a spatial random effect model involving both the spatial

[9] Samples that are considered outliers are identified based on variables values, which exhibit deviations exceeding two standard deviations from the expected values.

autoregressive terms for the spatial weight of GitHub working environment, $W^d$, and the geographical spatial weight, $W^s$[10].The mathematical form of the model is

$$\begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} \rho_1^d W_1^d & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho_k^d W_k^d \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} + \begin{bmatrix} \rho_1^s W_1^s & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho_k^s W_k^s \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} + \begin{bmatrix} X_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_k \end{bmatrix}.$$

In this equation, *k* is the country (regime) index. The first and the second components at the right hand side of the equation are the country-varying spatial lag terms. The third component is the country-varying covariates. This model posits that the work patterns of a focal user's online and offline neighbors correlate with that of the focal user and this correlation is heterogenous across different countries. This assumption is likely valid in the online context, as individuals who collaborate tend to work within the same time slots and similar frequency to ensure effective cooperation. However, it is uncertain how exactly physical neighbors bring influence on focal user. Thus, I also considered an alternative model specification as follows:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} \rho_1^d W_1^d & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho_k^d W_k^d \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} + \begin{bmatrix} X_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1^s \\ \vdots \\ \epsilon_k^s \end{bmatrix}$$

[10] I constructed $W^s$ as a Gaussian kernel weight, where the weight of each observation-pair was mathematically defined as $K(z_{ij}) = \sqrt{2\pi}e^{-z_{ij}^2/2}$ where $z_{ij} = \begin{cases} \frac{d_{ij}}{h_i} & d_{ij} \le h_i \\ 0 & d_{ij} > h_i \end{cases}$. $d_{ij}$ is the Euclidian distance between individual *i* and *j*. $h_i$ is the bandwidth for *i*, which was set as the max 6-nearest neighbors' distance of *i*. $W^d$ is transformed based on the GitHub users' collaboration network. In the collaboration network. Users that have directly interacted with the focal user were considered as the neighbors of that user and the corresponding position in the digital weight matrix was coded 1, otherwise 0.

where $\begin{bmatrix} \epsilon_1^s \\ \vdots \\ \epsilon_k^s \end{bmatrix} = \begin{bmatrix} \lambda_1^s W_1^s & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k^s W_k^s \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_k \end{bmatrix}$. Unlike the first model, this one suggests that the geographical spatial dependence is simply due to the geographic clustering of the distribution of the dependent variable. Table 4 presents the summarized regression results for the two model specifications, while a more detailed preliminary analysis and regression results can be found in Appendix B.

Table 4: Summarized spatial regression results

| | | **Two Spatial Lag** | | **Spatial Lag + Spatial Error** | |
|---|---|---|---|---|---|
| **Country** | **Dependent Variable** | $\rho^d$ | $\rho^s$ | $\rho^d$ | $\lambda^s$ |
| US | *Workday_Work_Intensity* | 0.216*** | 1.032** | 0.173*** | 0.031 |
| | | (0.061) | (0.331) | (0.010) | (0.137) |
| | *Weekend_Work_Intensity* | 0.184*** | 0.832** | 0.161*** | 0.115 |
| | | (0.011) | (0.359) | (0.040) | (0.124) |
| | *Workday_Weekend_Diff* | 0.246*** | 0.148 | 0.136*** | 0.134 |
| | | (0.013) | (0.164) | (0.020) | (0.108) |
| DE | *Workday_Work_Intensity* | 0.147*** | 0.959 | 0.131*** | -0.055 |
| | | (0.033) | (0.632) | (0.032) | (0.280) |
| | *Weekend_Work_Intensity* | 0.155*** | 4.170 | 0.173*** | 0.099 |
| | | (0.041) | (5.780) | (0.026) | (0.284) |
| | *Workday_Weekend_Diff* | 0.137*** | 0.754** | 0.181*** | 0.199 |
| | | (0.043) | (0.309) | (0.036) | (0.223) |
| CN | *Workday_Work_Intensity* | 0.193*** | 6.579 | 0.076*** | 0.387*** |
| | | (0.046) | (14.422) | (0.005) | (0.110) |
| | *Weekend_Work_Intensity* | 0.193*** | 1.820 | 0.099*** | -0.048 |
| | | (0.046) | (12.114) | (0.012) | (0.189) |
| | *Workday_Weekend_Diff* | 0.237*** | 0.775*** | 0.083*** | 0.428*** |
| | | (0.059) | (0.274) | (0.017) | (0.141) |
| IN | *Workday_Work_Intensity* | 0.135** | 2.704 | 0.107*** | 0.059 |
| | | (0.061) | (8.124) | (0.021) | (0.180) |
| | *Weekend_Work_Intensity* | 0.109 | 4.101 | 0.122*** | 0.300** |
| | | (0.099) | (6.052) | (0.028) | (0.136) |
| | *Workday_Weekend_Diff* | 0.088** | 2.313 | 0.114*** | -0.247 |
| | | (0.040) | (4.210) | (0.032) | (0.162) |

*Note: The three dependent variables are distinct metrics representing the working patterns of GitHub users. The Two Spatial Lag model incorporates both digital and geographical spatial lag elements, while the Spatial Lag + Spatial Error model combines the digital spatial lag with geographical spatial error elements.*

The three groups of regression analysis reveal notable differences in the coefficients of geographic space-related variables (i.e., $\rho^s$ and $\lambda^s$) and other covariates across the four countries, as confirmed by the Chow tests (see Appendix A). This heterogeneity reflects the different working cultures and open collaboration practices of different countries. However, $\rho^d$ remains significant in almost all cases, suggesting the universally robust impacts of local virtual workplace environments on GitHub users work patterns.

In the regression analysis of *Workday_Work_Intensity*, the spatial lag coefficient $\rho^s$ is only significant in the United States while the spatial error coefficient $\lambda^s$ is significant only in China, suggesting the business day work intensity of GitHub users in both the US and China exhibits noticeable spatial dependence, but in distinct ways. The working patterns in the US are influenced by those of their neighbors (spatial clustering of observed outcome) whereas the unobserved characteristics of workers in China are correlated with their neighbors that can explain the difference in work culture (spatial clustering of unobservable). Conversely, the business day work intensity of GitHub users in Germany and India does not demonstrate significant spatial dependence. When examining the regression analysis of *Weekend_Work_Intensity*, $\rho^s$ is significant only in the United States, while $\lambda^s$ is significant solely in India. This implies that the weekend work intensity of GitHub users in the US and India displays geographic spatial interdependence, but also in distinct manners. Finally, for *Workday_Weekend_Diff*, both $\rho^s$ and $\lambda^s$ are significant in China, and $\rho^s$ is significant in Germany, indicating that users' variation of work intensity between business days and weekends are geographically associated in the two countries. When examining the three dependent variables as indicators of distinct aspects of GitHub users' work patterns, it is noteworthy that while the variables associated with geographical spatial effects do not

consistently exhibit significance like the digital spatial effect indicator, we can still affirm that the work patterns of GitHub users in all four countries are influenced to varying degrees by their geographical distribution. Despite the diverse ways in which this influence manifests, it is evident that physical environment also plays a role in shaping the work patterns of GitHub users across the four countries. By far, the regression results confirm the dual medium effects on GitHub open collaboration participants' online dynamics.

# DISCUSSION

In this thesis, I introduce an innovative framework for exploring and modeling open collaboration. The foundation of this proposed framework is a loose analogy, considering individuals, teams, institutions, and social environments as akin to atoms, molecules, matter, and mediums in the physical realm. This methodology, however, does not immediately incorporate instruments from natural sciences into the investigation of human collaborative processes. Instead, it first highlights the importance of recognizing major disparities between analogous entities in different systems. This step is crucial for justifying the logic of model transfer. Following this, the framework adjusts and generalizes traditional natural science tools and concepts, thus enhancing their reliability for applications in social science research. As part of an empirical demonstration of this framework's functionality and the consolidation of the concepts proposed, it was employed to study the open collaboration practice on GitHub. Specifically, the fundamentals of the periodic table of elements, reaction equations, and the superposition state principle were adapted and applied to guide the investigation. These empirical evaluations made use of a large, independently collected dataset from GitHub. In the paragraphs that follow, I will delve into the various discoveries and insights that this study has brought to light thus far.

First, this study uncovers novel empirical findings, enhancing our understanding of the intricate dynamics of open collaboration practices on GitHub. On an individual level, the analysis identifies six latent categories of participants within the community. These categories emerge from a consideration of personal background information and collaborative behaviors. The application of dynMMSBM regression further delineates the

unique collaborative habits inherent to these six user categories. The collaboration periodic table offers a more nuanced perspective, accentuating disparities in the distribution of programming expertise and activity levels among these categories. Moreover, it provides a graphical representation enabling evaluation of an individual's potential category affiliations. It's revealed that the majority of individuals defy straightforward classification into a single category. Even those who can be somewhat confidently assigned to a specific category exhibit substantial fluctuations in their category affiliations between 2022 and 2023. To further illuminate the uncertainty and temporal shifts observed in GitHub users' mixed category representations, regime spatial econometric models were applied. The analysis posits that the work patterns of GitHub users are simultaneously shaped by various online and offline environments they inhabit. This interplay imbues their behavior with a level of unpredictability, and this unpredictability was formally introduced as the ‘social superposition state’ in this research. On a group level, five distinct types of open collaboration structures in GitHub projects were discerned. The construction of probabilistic collaboration equations enabled the identification of common types of GitHub users involved in different types of open collaboration structures. These findings illuminate the relationship between individual users’ characteristics and the coordination strategies of their respective open collaboration groups.

Second, this research, standing as one of the pioneering attempts to systematically transfer methods typically used for understanding physical particle interaction to model open collaboration, showcases the potential for utilizing this strategy in the study of a wider array of human interaction scenarios as well. At the onset of this paper, I postulated that this analogical approach would be particularly effective when studying open collaboration.

Digital platforms provide collaborators with greater autonomy in deciding their collaborative engagement, a process that more closely mirrors the interaction of particles. Nevertheless, this does not preclude the applicability of this analogous framework to the study of other forms of human interactions, such as traditional organizational collaboration. Drawing parallels with the natural world, just as physical particles modify their reaction modalities and exhibited properties based on their environment, individuals within a social construct also adapt their behaviors in response to their surroundings. Traditional collaboration predominantly occurs within physically confined workspaces, which can significantly influence people's interaction patterns and personal dynamics. Investigating these differences within traditional workspaces and digital workspaces can provide valuable insights into the influence of medium factors on shaping interaction patterns. Moreover, the efforts on extending this analogical framework to other scenarios can facilitate the adaptation and application of a broader array of physicochemical concepts and instruments to model human interactions. This ultimately contributes to the completion and expansion of the proposed framework as a comprehensive strategy for modeling and investigating human interaction, extending its utility beyond the realm of open collaboration.

Third, from a more audacious standpoint, this study manages to carve out a new trajectory for future progress in sociophysics. It does so by presenting a strategy that may be less malleable, but is indeed more systematic and reliable, and by providing tangible evidence of its practical utility. Previous attempts to transpose methods from physics for the study of social sciences have often overlooked justifying the fundamental analogical assumptions. Besides, much of the related research heavily leans on simulation methods, which often leads to the proposed models lacking in empirical validation. Furthermore, critics from the realm

of social science frequently argue that these transplanted models from physics tend to drastically oversimplify their target social systems. This leads to what can be perceived as an almost ostentatious game of dress-up, straining the analogy to its limits while failing to provide a genuine and nuanced understanding of the social phenomena in question. To address these issues, the strategy proposed in this study is centered around constructing a bottom-up social-physical analogy. This approach involves loosely aligning and contrasting core elements from both physical and social systems. Based on a rigorous discussion on the similarities and discrepancies between each analogical element-pair, the concepts and tools initially developed in the natural sciences are adjusted to more accurately model social systems. These adapted concepts and tools are then validated and further refined based on real-world data observed in various social contexts. This strategy ensures a more robust and empirically grounded exploration of social phenomena through the lens of physical sciences. Over time, it has the potential to evolve from a mere discussion on cross-paradigm methodological transfer to laying the foundation for a fully-fledged theory of social physics. This strategy could, in effect, bridge the gap between physical and social sciences, offering a holistic, integrated perspective on human behavior and interactions.

Despite the many contributions of this study, some limitations must be acknowledged, which also provide directions for future research. The first limitation of the present study lies in the limited computing resources. This study was conducted using a personal MacBook laptop equipped with a Mac1 chip for all computational tasks. Given the considerable data volume and the computational demand imposed by dynMMSBM, a limited sampling strategy was adopted. This encompassed only 10% of the available dataset for the regression analysis and the subsequent construction of the collaboration periodic table. Additionally, due to

memory limitations, the estimation of the covariance matrix was not feasible; therefore, only the regression variable coefficients were reported. While these constraints may affect the comprehensiveness of the findings, they represent necessary compromises given the current resources. Moving forward, more robust insights can potentially be gleaned from implementing the collaboration periodic table using more advanced computational resources, such as high-performance computing clusters, which would allow for a more extensive data sampling and a complete estimation of all statistical parameters.

The second limitation concerns the interpretability of the current version of the collaboration periodic table. By comparison, chemists have devoted substantial time and effort into developing and refining the periodic table of elements, transforming it into a potent tool for interpretation and prediction. In contrast, the collaboration periodic table proposed in this study represents a pioneering attempt and, as such, has a number of apparent shortcomings. The current challenges associated with extracting meaningful information from the collaboration periodic table should not, however, diminish its potential value. To enhance the efficacy and applicability of the social science version of the periodic table to match that of its chemical counterpart, future research endeavors could focus on expanding the current construction approach or propose novel methodologies. This would allow the periodic table to evolve and improve over time, gaining in depth, breadth, and sophistication.

The third limitation centers around the rigidity of the spatial regression analysis used for assessing social superposition states. In this study, I treated the digital spatial weight as an exogenous variable, on par with geographical weight. It's crucial to note, however, that a GitHub user's local collaboration network could be significantly impacted by their work habits, suggesting an endogenous nature to the digital spatial weight. Unfortunately, due to

the absence of an open-source algorithm to handle endogenous weight issues, a simplified approach was employed. As we move forward, future research could tackle the issue of endogenous weight with greater precision and consideration. This could potentially validate the robustness of the conclusions drawn from this segment of the research and contribute to a more nuanced understanding of the complexities involved.

Finally, it is important to note that while this study has been tailored to the specific context of open collaboration, the framework proposed herein has broader implications. Future research could apply and, more importantly, extend this framework to a variety of other social contexts characterized by complex human interactions—scenarios that traditional social science tools may find challenging to thoroughly investigate. With the substantial volume of social data available today, it is now feasible to develop a suite of new sociophysical instruments from scratch, thereby fostering the evolution of a comprehensive sociophysical theory in the foreseeable future. This represents a rich and fertile area for future explorations and advancements, with the potential to significantly impact our understanding of complex social interactions and the factors that influence them.

# REFERENCES


Abergel, F., Aoyama, 1954-, Hideaki, Chakrabarti, 1952-, B. K. (Bikas K. )., Chakraborti, A., Deo, N., Raina, D., & Vodenska, I. (Eds.). (2017). *Econophysics and sociophysics: Recent progress and future directions*. Springer. https://link.springer.com/10.1007/978-3-319-47705-3

Arnopoulos, P. (2005). *Sociophysics: Cosmos and chaos in nature and culture*. Nova Science Publishers.

Avieson, B. (2022). Editors, sources and the'go back'button: Wikipedia's framework for beating misinformation. *First Monday*.

Baldwin, C., & von Hippel, E. (2011). Modeling a Paradigm Shift: From Producer Innovation to User and Open Collaborative Innovation. *Organization Science*, *22*(6), 1399–1417. https://doi.org/10.1287/orsc.1100.0618

Berlingerio, M., Koutra, D., Eliassi-Rad, T., & Faloutsos, C. (2012). NetSimile: A Scalable Approach to Size-Independent Network Similarity. *CoRR*, *abs/1209.2684*. http://arxiv.org/abs/1209.2684

Bernstein, E. S., & Turban, S. (2018). The impact of the 'open' workspace on human collaboration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1753), 20170239. https://doi.org/10.1098/rstb.2017.0239

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Botsman, R., & Rogers, R. (2010). *What's Mine Is Yours: The Rise of Collaborative Consumption*. HarperCollins. https://books.google.com/books?id=LiC2foFeXQYC

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, *81*(2), 591–646. https://doi.org/10.1103/RevModPhys.81.591

Cho, A. (2009). Ourselves and Our Interactions: The Ultimate Physics Problem? *Science*, *325*(5939), 406–408. https://doi.org/10.1126/science.325_406

David, P. A. (1998). Common Agency Contracting and the Emergence of "Open Science" Institutions. *The American Economic Review*, *88*(2), 15–21. JSTOR.

Dennison, J. (2019). A review of public issue salience: Concepts, determinants and effects on voting. *Political Studies Review*, *17*(4), 436–446.

Du Chatenier, E., Verstegen, J. A., Biemans, H. J., Mulder, M., & Omta, O. (2009). The challenges of collaborative knowledge creation in open innovation teams. *Human Resource Development Review*, *8*(3), 350–381.

Eklund, L., Stamm, I., & Liebermann, W. K. (2019). The crowd in crowdsourcing: Crowdsourcing as a pragmatic research method. *First Monday*, *4*(10). https://doi.org/10.5210/fm.v24i10.9206

Ferraro, F., & O'Mahony, S. (2012). 545Managing the Boundaries of an "Open" Project. In J. F. Padgett & W. W. Powell (Eds.), *The Emergence of Organizations and Markets* (p. 0). Princeton University Press. https://doi.org/10.23943/princeton/9780691148670.003.0018

Forte, A., & Lampe, C. (2013). Defining, Understanding, and Supporting Open Collaboration: Lessons From the Literature. *American Behavioral Scientist*, *57*(5), 535–547. https://doi.org/10.1177/0002764212469362

Foucault, M. (1966). *The Order of Things*.

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy*, *43*(1), 1–20.

Galam, S. (2012). *Sociophysics: A physicist's modeling of psycho-political phenomena*. Springer Verlag.

Galam, S. (2013). Modeling the Forming of Public Opinion: An approach from Sociophysics. *Global Economics and Management Review*, *18*(1), 2–11. https://doi.org/10.1016/S2340-1540(13)70002-1

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*(7070), 900–901. https://doi.org/10.1038/438900a

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380. JSTOR.

Humphreys, P. (2019). Knowledge transfer across scientific disciplines. *Studies in History and Philosophy of Science Part A*, *77*, 112–119. https://doi.org/10.1016/j.shpsa.2017.11.001

Ishii, A., & Okano, N. (2021). Sociophysics Approach of Simulation of Mass Media Effects in Society Using New Opinion Dynamics. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Intelligent Systems and Applications* (pp. 13–28). Springer International Publishing.

Jensen, P. (2019). The politics of physicists' social models. *Comptes Rendus Physique*, *20*(4), 380–386. https://doi.org/10.1016/j.crhy.2019.05.016

Jusup, M., Holme, P., Kanazawa, K., Takayasu, M., Romić, I., Wang, Z., Geček, S., Lipić, T., Podobnik, B., Wang, L., Luo, W., Klanjšček, T., Fan, J., Boccaletti, S., & Perc, M.

(2022). Social physics. *Social Physics*, *948*, 1–148. https://doi.org/10.1016/j.physrep.2021.10.005

Kuhn, T. S., & Hacking, I. (2012). *The Structure of Scientific Revolutions*. University of Chicago Press. https://books.google.com/books?id=3eP5Y_OOuzwC

Levine, S. S., & Prietula, M. J. (2014). Open Collaboration for Innovation: Principles and Performance. *Organization Science*, *25*(5), 1414–1433. https://doi.org/10.1287/orsc.2013.0872

Lin, C.-H. (2022). Knowledge transfer, templates, and the spillovers. *European Journal for Philosophy of Science*, *12*(1), 6. https://doi.org/10.1007/s13194-021-00426-w

Mergel, I. (2015). Open collaboration in the public sector: The case of social coding on GitHub. *Government Information Quarterly*, *32*(4), 464–472. https://doi.org/10.1016/j.giq.2015.09.004

Olivella, S., Pratt, T., & Imai, K. (2022). Dynamic Stochastic Blockmodel Regression for Network Data: Application to International Militarized Conflicts. *Journal of the American Statistical Association*, *117*(539), 1068–1081. https://doi.org/10.1080/01621459.2021.2024436

O'Mahony, S., & Bechky, B. A. (2008). Boundary Organizations: Enabling Collaboration among Unexpected Allies. *Administrative Science Quarterly*, *53*(3), 422–459. https://doi.org/10.2189/asqu.53.3.422

Parongama Sen & Bikas K. Chakrabarti. (2014). *Sociophysics: An Introduction: Vol. First edition*. OUP Oxford; eBook Academic Collection (EBSCOhost). http://proxy.uchicago.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=e000xna&AN=643991&site=eds-live&scope=site

Pentland, 1952-, Alex. (2015). *Social physics: How social networks can make us smarter*. Penguin Books.

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, ou, Essai de physique sociale*. Bachelier. https://books.google.com/books?id=VXkZAAAAYAAJ

Riehle, D., Ellenberger, J., Menahem, T., Mikhailovski, B., Natchetoi, Y., Naveh, B., & Odenwald, T. (2009). Open Collaboration within Corporations Using Software Forges. *IEEE Software*, *26*(2), 52–58. https://doi.org/10.1109/MS.2009.44

Rob McLean. (2020, June 25). These companies plan to make working from home the new normal. As in forever. *CNN Business*. https://www.cnn.com/2020/05/22/tech/work-from-home-companies/index.html

Scaliza, J. A. A., Jugend, D., Chiappetta Jabbour, C. J., Latan, H., Armellini, F., Twigg, D., & Andrade, D. F. (2022). Relationships among organizational culture, open innovation, innovative ecosystems, and performance of firms: Evidence from an emerging economy

context. *Journal of Business Research*, *140*, 264–279. https://doi.org/10.1016/j.jbusres.2021.10.065

Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2021). Socio-Spatial Properties of Online Location-Based Social Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 329–336. https://doi.org/10.1609/icwsm.v5i1.14094

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, *1*(2), 143–186. https://doi.org/10.1080/0022250X.1971.9989794

Schweitzer, F. (2018). Sociophysics. *Phys. Today*, *71*(2), 40.

Splitter, V., Dobusch, L., Krogh, G. von, Whittington, R., & Walgenbach, P. (2023). Openness as Organizing Principle: Introduction to the Special Issue. *Organization Studies*, *44*(1), 7–27. https://doi.org/10.1177/01708406221145595

Stauffer, D. (2002). Sociophysics: The Sznajd model and its applications. *Proceedings of the STATPHYS Satellite Conference: Challenges in Computational Statistical Physics in Teh 21st Century*, *146*(1), 93–98. https://doi.org/10.1016/S0010-4655(02)00439-3

Tan, P. (2023). Interdisciplinary model transfer and realism about physical analogy. *Synthese*, *201*(2), 65. https://doi.org/10.1007/s11229-023-04065-x

Vazquez, F. (2022). Modeling and analysis of social phenomena: Challenges and possible research directions. In *Entropy* (Vol. 24, Issue 4, p. 491). MDPI.

Wang, T., Wu, J., Gu, J., & Hu, L. (2021). Impact of open innovation on organizational performance in different conflict management styles: Based on resource dependence theory. *International Journal of Conflict Management*, *32*(2), 199–222.

Yang, L., Holtz, D., Jaffe, S., Suri, S., Sinha, S., Weston, J., Joyce, C., Shah, N., Sherman, K., Hecht, B., & Teevan, J. (2022). The effects of remote work on collaboration among information workers. *Nature Human Behaviour*, *6*(1), 43–54. https://doi.org/10.1038/s41562-021-01196-4

APPENDIX A

Table A1: Topics derived from LDA and their top-10 keywords ($\lambda = 0.5$)

| Topic | Keywords |
|---|---|
| *Programming Language* | config, github-config, typescript, react, hacktoberfest, javascript, golang, android, go, linux |
| *Data Science* | deep-learning, pytorch, machine-learning, computer-vision, tailwindcss, laravel, artificial-intelligence, nlp, object-detection, ai |
| *Software Engineering* | flutter, swift, neovim, ios, discord, swiftui, database, bot, dart, lua |
| *Information Security* | security, python, docker, python3, hacking, security-tools, bugbounty, svelte, pentesting, cybersecurity |

Table A2: Top 3 association rules for each social molecule prototype

| Prototype | Antecedents | Support | Confidence | Lift |
|---|---|---|---|---|
| *Parallel* | IV-H | 0.012 | 0.304 | 1.780 |
| | IV-C, IV-D, VI-C | 0.011 | 0.259 | 1.519 |
| | IV-D, VI-C | 0.015 | 0.238 | 1.392 |
| *Complex* | III-C, III-D | 0.011 | 0.075 | 1.388 |
| | III-C | 0.020 | 0.065 | 1.197 |
| | V-D | 0.015 | 0.063 | 1.164 |
| *Structural* | II-D, III-E, IV-D | 0.010 | 0.488 | 1.991 |
| | II-C, III-E | 0.013 | 0.481 | 1.966 |
| | II-D, III-D, III-E | 0.011 | 0.429 | 1.750 |
| *Linear* | VI-H | 0.012 | 0.411 | 1.484 |
| | II-D, III-C, V-C | 0.020 | 0.371 | 1.342 |
| | III-C, IV-C, IV-E | 0.011 | 0.361 | 1.303 |
| *Star* | IV-D, IV-E, V-C | 0.011 | 0.438 | 1.726 |
| | III-F, IV-D | 0.010 | 0.426 | 1.678 |
| | IV-D, IV-F | 0.014 | 0.360 | 1.420 |

Table A3: Descriptive statistics of the sub-sample sets

| Country | Variables | Mean | Std | Min | Max |
|---|---|---|---|---|---|
| United States | *Workday_Work_Intensity* | 3.788 | 7.286 | 0.000 | 64.750 |
| | *Weekend_Work_Intensity* | 1.925 | 4.151 | 0.000 | 76.000 |
| | *Difference_Work_Intensity* | 1.863 | 6.347 | -18.38 | 57.083 |
| | *Experience* | 8.632 | 3.501 | 0.016 | 15.126 |
| | *Productivity* | 1.766 | 3.095 | 0.000 | 33.144 |
| | *Skill_Diversity* | 2.731 | 0.777 | 0.000 | 5.313 |
| | *Engagement* | 32.73 | 264.0 | 0.000 | 11928 |

| | | | | | |
|---|---|---|---|---|---|
| | *Reputation* | 160.2 | 1116 | 0.000 | 38916 |
| | *Freelancer* | 0.155 | 0.362 | 0.000 | 1.000 |
| China | *Workday_Work_Intensity* | 2.343 | 5.264 | 0.000 | 72.000 |
| | *Weekend_Work_Intensity* | 1.502 | 3.055 | 0.000 | 59.500 |
| | *Difference_Work_Intensity* | 0.840 | 4.738 | -17.22 | 35.500 |
| | *Experience* | 6.428 | 2.698 | 0.016 | 14.890 |
| | *Productivity* | 1.248 | 2.402 | 0.000 | 56.131 |
| | *Skill_Diversity* | 2.807 | 0.751 | 0.000 | 4.942 |
| | *Engagement* | 46.74 | 347.2 | 0.000 | 24175 |
| | *Reputation* | 156.3 | 1413.6 | 0.000 | 74064 |
| | *Freelancer* | 0.207 | 0.405 | 0.000 | 1.000 |
| Germany | *Workday_Work_Intensity* | 3.712 | 7.148 | 0.000 | 97.000 |
| | *Weekend_Work_Intensity* | 2.009 | 4.092 | 0.000 | 72.333 |
| | *Difference_Work_Intensity* | 1.703 | 6.377 | -19.77 | 49.333 |
| | *Experience* | 8.187 | 3.360 | 0.008 | 15.016 |
| | *Productivity* | 4.855 | 5.800 | 0.000 | 49.194 |
| | *Skill_Diversity* | 2.688 | 0.850 | 0.000 | 5.193 |
| | *Engagement* | 23.80 | 61.21 | 0.000 | 2033.0 |
| | *Reputation* | 80.73 | 418.8 | 0.000 | 17709 |
| | *Freelancer* | 0.192 | 0.394 | 0.000 | 1.000 |
| India | *Workday_Work_Intensity* | 2.095 | 3.615 | 0.000 | 56.800 |
| | *Weekend_Work_Intensity* | 1.408 | 2.570 | 0.000 | 31.000 |
| | *Difference_Work_Intensity* | 0.687 | 3.088 | -18.94 | 51.800 |
| | *Experience* | 5.341 | 2.958 | 0.605 | 15.002 |
| | *Productivity* | 8.594 | 8.882 | 0.000 | 49.353 |
| | *Skill_Diversity* | 2.620 | 0.762 | 0.000 | 5.106 |
| | *Engagement* | 27.20 | 108.8 | 0.000 | 2960.0 |
| | *Reputation* | 64.11 | 379.8 | 0.000 | 12075 |
| | *Freelancer* | 0.331 | 0.471 | 0.000 | 1.000 |

# APPENDIX B

In this appendix, I provide a more detailed explanation of the spatial econometric analysis carried out in Section 5. As an initial check of the spatial effect, I visualized the spatial distributions of the three interested variables, *Workday_Work_Intensity*, *Weekend_Work_Intensity*, and *Difference_Work_Intensity* using spatial quantile maps (i.e., Figure B1, B2, and B3).

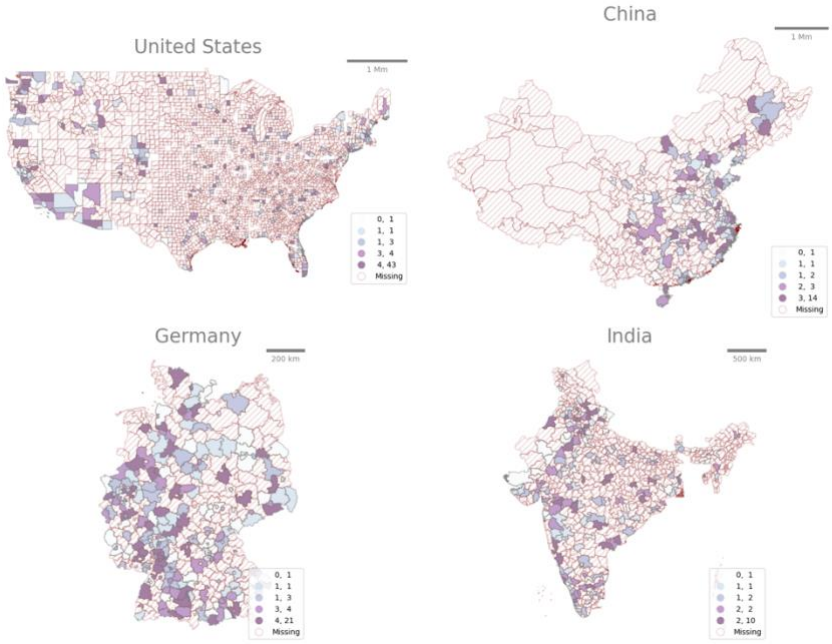


Figure B1: The spatial distribution of *Workday_Work_Intensity*

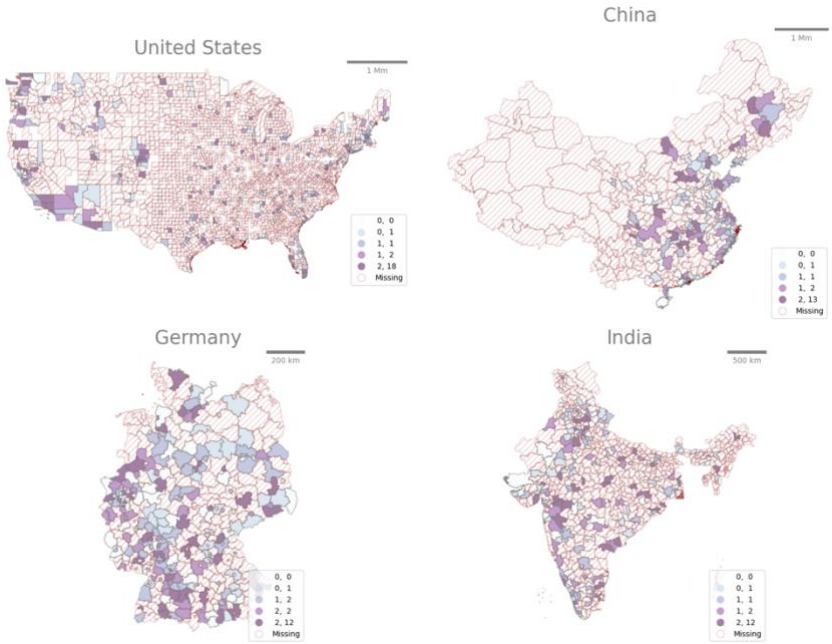


Figure B2: The spatial distribution of *Weekend_Work_Intensity*

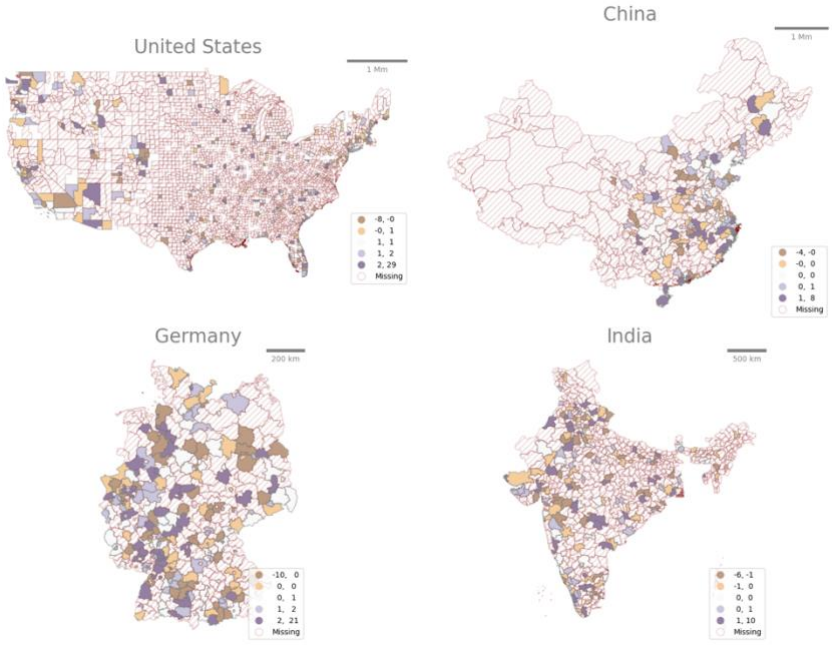


Figure B3: The spatial distribution of *Difference_Work_Intensity*

I calculated the Moran' Is of the three dependent variables and assessed their significance by performing a permutation test (9999 permutations). The results are reported in Table B1.

Table B1: Univariate Moran's I statistics and permutation test

| **Country** | **United States** | **China** | **Germany** | **India** |
|---|---|---|---|---|
| *Workday_Work_Intensity* | 0.003** | 0.004*** | 0.002* | 0.003 |
| | (0.044) | (0.002) | (0.075) | (0.144) |
| *Weekend_Work_Intensity* | 0.002* | 0.001 | 0.000 | 0.013*** |
| | (0.086) | (0.134) | (0.485) | (0.003) |
| *Difference_Work_Intensity* | 0.005*** | 0.003*** | 0.001 | -0.001 |
| | (0.009) | (0.005) | (0.116) | (0.407) |

*Note.* Pseudo p-value are shown in the parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The analysis of the figures and the univariate Moran's I permutation test reveals the presence of spatial autocorrelation in the work patterns of GitHub users among the four countries. To strengthen the evidence of spatial effects, I then conducted OLS regression as well as employed the two spatial models described in section 5 using the dataset. Table B2, B3, and B4 report the regression results considering *Workday_Work_Intensity*, *Weekend_Work_Intensity*, and *Difference_Work_Intensity*, respectively, as dependent variables.

Table B2: Regression results for *Workday_Work_Intensity*

| | United States | | | China | | | Germany | | | India | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **OLS** | **M1** | **M2** | **OLS** | **M1** | **M2** | **OLS** | **M1** | **M2** | **OLS** | **M1** | **M2** |
| *Experience* | 0.093*** | -0.585*** | 0.093*** | 0.069*** | 1.276*** | 0.064** | 0.078** | 0.137 | 0.053 | 0.089*** | 0.128*** | 0.089*** |
| | (0.035) | (0.124) | (0.035) | (0.026) | (0.034) | (0.031) | (0.038) | (0.126) | (0.036) | (0.030) | (0.037) | (0.029) |
| *Productivity* | 0.069*** | 0.051*** | 0.069*** | 0.028*** | 0.067*** | 0.026*** | 0.302*** | 0.370*** | 0.272*** | 0.056*** | 0.030** | 0.056*** |
| | (0.009) | (0.035) | (0.010) | (0.005) | (0.007) | (0.005) | (0.021) | (0.144) | (0.039) | (0.010) | (0.013) | (0.010) |
| *Skill_Diversity* | 0.737*** | 3.541*** | 0.736*** | 0.548*** | -1.631*** | 0.549*** | 0.643*** | -1.154 | 0.529*** | 0.056 | 0.440*** | 0.056 |
| | (0.165) | (0.550) | (0.165) | (0.098) | (0.137) | (0.098) | (0.157) | (2.479) | (0.154) | (0.115) | (0.137) | (0.115) |
| *Engagement* | 0.0002 | 0.0001 | 0.0001 | 0.002*** | 0.002*** | 0.002*** | 0.0003 | 0.000 | 0.001 | 0.002*** | 0.003*** | 0.002*** |
| | (0.0004) | (0.001) | (0.0004) | (0.000) | (0.000) | (0.000) | (0.002) | (0.002) | (0.002) | (0.001) | (0.000) | (0.001) |
| *Reputation* | 0.001*** | 0.001 | 0.001** | 0.0001*** | 0.0001** | 0.0001*** | 0.002*** | 0.002*** | 0.002** | 0.001*** | 0.001** | 0.001*** |
| | (0.000) | (0.000) | (0.000) | (0.0000) | (0.0001) | (0.0000) | (0.000) | (0.000) | (0.001) | (0.000) | (0.000) | (0.000) |
| *Freelancer* | -0.760*** | -1.578 | -0.761*** | 0.483*** | -2.716*** | 0.389** | -0.997*** | -2.266 | -0.934*** | -0.049 | 0.877*** | -0.048 |
| | (0.271) | (1.044) | (0.270) | (0.184) | (0.229) | (0.181) | (0.290) | (1.866) | (0.269) | (0.164) | (0.240) | (0.163) |
| $\rho^d$ | | 0.216*** | 0.173*** | | 0.193*** | 0.076*** | | 0.147*** | 0.131*** | | 0.135** | 0.107*** |
| | | (0.061) | (0.010) | | (0.046) | (0.005) | | (0.033) | (0.032) | | (0.061) | (0.021) |
| $\rho^s$ | | 1.032** | | | 6.579 | | | 0.959 | | | 2.704 | |
| | | (0.331) | | | (14.422) | | | (0.632) | | | (8.124) | |
| $\lambda^s$ | | | 0.031 | | | 0.387*** | | | -0.055 | | | 0.059 |
| | | | (0.137) | | | (0.110) | | | (0.280) | | | (0.180) |
| *Main_Language* | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| *Institution* | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| **Tests** | Statistics | | p-value | Statistics | | p-value | Statistics | | p-value | Statistics | | p-value |
| Rob-LM (lag) | 4.588 | | 0.032 | 7.319 | | 0.007 | 2.120 | | 0.145 | 1.182 | | 0.670 |
| Rob-LM (error) | 4.742 | | 0.029 | 6.682 | | 0.010 | 1.248 | | 0.264 | 1.169 | | 0.681 |
| **Global Tests** | Statistics | | | | | | p-value | | | | | |
| Rob-LM (lag) | 5.824 | | | | | | 0.016 | | | | | |
| Rob-LM (error) | 6.052 | | | | | | 0.014 | | | | | |
| Chow test (OLS) | 650.044 | | | | | | 0.000 | | | | | |
| Chow test (M1) | 24869.048 | | | | | | 0.000 | | | | | |
| Chow test (M2) | 656.129 | | | | | | 0.000 | | | | | |

*Note.* Standard errors are shown in the parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table B3. Regression results for *Weekend_Work_Intensity*

| | United States | | | China | | | Germany | | | India | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **OLS** | **M1** | **M2** | **OLS** | **M1** | **M2** | **OLS** | **M1** | **M2** | **OLS** | **M1** | **M2** |
| *Experience* | 0.064*** | 0.060*** | 0.061** | 0.069*** | 0.300** | 0.064** | 0.009 | 0.066 | 0.003 | -0.007 | -0.050 | -0.004 |
| | (0.021) | (0.020) | (0.020) | (0.026) | (0.129) | (0.031) | (0.022) | (0.108) | (0.021) | (0.021) | (0.134) | (0.020) |
| *Productivity* | 0.041*** | 0.039*** | 0.039*** | 0.028*** | 0.023*** | 0.023*** | 0.162*** | 0.211* | 0.147*** | 0.058*** | -0.005 | 0.053*** |
| | (0.005) | (0.005) | (0.012) | (0.005) | (0.003) | (0.005) | (0.012) | (0.113) | (0.022) | (0.007) | (0.117) | (0.008) |
| *Skill_Diversity* | 0.542*** | 0.396*** | 0.394*** | 0.661*** | 0.342*** | 0.552*** | 0.352*** | -0.875 | 0.276*** | 0.086 | 1.319 | 0.022 |
| | (0.098) | (0.096) | (0.117) | (0.097) | (0.057) | (0.077) | (0.091) | (2.015) | (0.084) | (0.082) | (2.525) | (0.094) |
| *Engagement* | 0.0005** | 0.0005*** | 0.0005 | 0.002*** | 0.002*** | 0.002*** | 0.001 | 0.001 | 0.001 | 0.002*** | 0.001 | 0.002 |
| | (0.0002) | (0.0002) | (0.0004) | (0.000) | (0.000) | (0.000) | (0.001) | (0.002) | (0.001) | (0.001) | (0.002) | (0.001) |
| *Reputation* | 0.0002*** | 0.0002*** | 0.0002** | 0.0001*** | 0.0001*** | 0.0001* | 0.001*** | 0.001*** | 0.001*** | 0.0004*** | 0.0003 | 0.0004 |
| | (0.0000) | (0.0000) | (0.0001) | (0.0000) | (0.0000) | (0.0000) | (0.000) | (0.000) | (0.000) | (0.0001) | (0.0003) | (0.0003) |
| *Freelancer* | -0.027 | 0.010 | 0.008 | 0.483*** | 0.449*** | 0.450*** | -0.242 | -1.064 | -0.192 | 0.113 | 1.260 | 0.155 |
| | (0.161) | (0.157) | (0.149) | (0.107) | (0.105) | (0.132) | (0.168) | (1.500) | (0.159) | (0.119) | (2.525) | (0.128) |
| $\rho^d$ | | 0.184*** | 0.161*** | | 0.193*** | 0.099*** | | 0.155*** | 0.173*** | | 0.109 | 0.122*** |
| | | (0.011) | (0.040) | | (0.046) | (0.012) | | (0.041) | (0.026) | | (0.099) | (0.028) |
| $\rho^s$ | | 0.832** | | | 1.820 | | | 4.170 | | | 4.101 | |
| | | (0.359) | | | (12.114) | | | (5.780) | | | (6.052) | |
| $\lambda^s$ | | | 0.115 | | | -0.048 | | | 0.099 | | | 0.300** |
| | | | (0.124) | | | (0.189) | | | (0.284) | | | (0.136) |
| *Main_Language* | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| *Institution* | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| **Tests** | Statistics | | p-value | Statistics | | p-value | Statistics | | p-value | Statistics | | p-value |
| Rob-LM (lag) | 2.486 | | 0.115 | 0.000 | | 0.999 | 0.057 | | 0.811 | 1.665 | | 0.197 |
| Rob-LM (error) | 1.572 | | 0.210 | 0.019 | | 0.889 | 0.033 | | 0.857 | 0.234 | | 0.628 |
| **Global Tests** | Statistics | | | | | | p-value | | | | | |
| Rob-LM (lag) | 0.862 | | | | | | 0.353 | | | | | |
| Rob-LM (error) | 0.255 | | | | | | 0.613 | | | | | |
| Chow test (OLS) | 407.470 | | | | | | 0.000 | | | | | |
| Chow test (M1) | 84.022 | | | | | | 0.000 | | | | | |
| Chow test (M2) | 389.602 | | | | | | 0.000 | | | | | |

*Note.* Standard errors are shown in the parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table B4: Regression results for *Difference_Work_Intensity*

| | United States | | | China | | | Germany | | | India | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **OLS** | **M1** | **M2** | **OLS** | **M1** | **M2** | **OLS** | **M1** | **M2** | **OLS** | **M1** | **M2** |
| *Experience* | 0.031 | 0.031 | 0.032 | 0.067*** | 0.063*** | 0.062* | 0.069 | 0.089 | 0.047 | 0.105*** | 0.353 | 0.096*** |
| | (0.031) | (0.030) | (0.032) | (0.025) | (0.024) | (0.033) | (0.035) | (0.097) | (0.031) | (0.026) | (0.610) | (0.027) |
| *Productivity* | 0.033*** | 0.292*** | 0.029* | 0.003 | 0.002 | 0.002 | 0.141*** | 0.171 | 0.121*** | 0.005 | 0.069 | 0.001 |
| | (0.008) | (0.008) | (0.017) | (0.005) | (0.004) | (0.003) | (0.020) | (0.108) | (0.036) | (0.009) | (0.160) | (0.010) |
| *Skill_Diversity* | 0.535*** | 0.279** | 0.278 | 0.242*** | 0.189** | 0.181** | 0.291** | -0.545 | 0.243* | 0.043 | -1.318 | 0.006 |
| | (0.150) | (0.143) | (0.178) | (0.091) | (0.089) | (0.070) | (0.146) | (1.851) | (0.137) | (0.100) | (3.181) | (0.100) |
| *Engagement* | -0.0003 | -0.0004 | -0.0004 | 0.002*** | 0.002*** | 0.002*** | -0.001 | -0.001 | -0.001 | 0.0002 | 0.000 | 0.000 |
| | (0.0004) | (0.0004) | (0.0003) | (0.000) | (0.000) | (0.000) | (0.002) | (0.002) | (0.002) | (0.0007) | (0.001) | (0.001) |
| *Reputation* | 0.0005*** | 0.0005*** | 0.0005* | 3.91e-5 | 4.16e-5 | 4.15e-5 | 0.001** | 0.001 | 0.001 | 0.0002 | 0.000 | 0.0003 |
| | (0.0000) | (0.0000) | (0.0003) | (4.16e-5) | (4.09e-5) | (3.65e-5) | (0.000) | (0.001) | (0.001) | (0.0002) | (0.001) | (0.0002) |
| *Freelancer* | -0.897*** | -0.714*** | -0.717*** | -0.055 | -0.064 | -0.065 | -0.754*** | -1.388 | -0.752*** | -0.226 | 0.080 | -0.225 |
| | (0.247) | (0.236) | (0.199) | (0.068) | (0.165) | (0.166) | (0.269) | (1.366) | (0.256) | (0.145) | (0.780) | (0.144) |
| $\rho^d$ | | 0.246*** | 0.136*** | | 0.237*** | 0.083*** | | 0.137*** | 0.181*** | | 0.088** | 0.114*** |
| | | (0.013) | (0.020) | | (0.059) | (0.017) | | (0.043) | (0.036) | | (0.040) | (0.032) |
| $\rho^s$ | | 0.148 | | | 0.775*** | | | 0.754** | | | 2.313 | |
| | | (0.164) | | | (0.274) | | | (0.309) | | | (4.210) | |
| $\lambda^s$ | | | 0.134 | | | 0.428*** | | | 0.199 | | | -0.247 |
| | | | (0.108) | | | (0.141) | | | (0.223) | | | (0.162) |
| *Main_Language* | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| *Institution* | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| **Tests** | Statistics | | p-value | Statistics | | p-value | | | p-value | Statistics | | p-value |
| Rob-LM (lag) | 6.946 | | 0.008 | 3.913 | | 0.048 | 2.669 | | 0.102 | 0.356 | | 0.551 |
| Rob-LM (error) | 4.894 | | 0.027 | 2.496 | | 0.114 | 1.477 | | 0.224 | 0.792 | | 0.374 |
| **Global Tests** | Statistics | | | | | | p-value | | | | | |
| Rob-LM (lag) | 11.232 | | | | | | 0.001 | | | | | |
| Rob-LM (error) | 6.193 | | | | | | 0.013 | | | | | |
| Chow test (OLS) | 281.335 | | | | | | 0.000 | | | | | |
| Chow test (M1) | 64.663 | | | | | | 0.055 | | | | | |
| Chow test (M2) | 525.927 | | | | | | 0.000 | | | | | |

*Note.* Standard errors are shown in the parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Intriguingly, the three groups of regression analysis reveal notable differences in the coefficients of geographic space-related variables (i.e., $\rho^s$ and $\lambda^s$) and other covariates across the four countries, as confirmed by the Chow tests. This heterogeneity reflects the different working cultures and open collaboration practices of different countries. However, among all the cases, $\rho^d$ remains significant, suggesting the universally robust impacts of local virtual workplace environments on GitHub users work patterns.

In the regression analysis of *Workday_Work_Intensity*, the spatial lag coefficient $\rho^s$ is only significant in the United States while the spatial error coefficient $\lambda^s$ is significant only in China. This suggests that the business day work intensity of GitHub users in both the US and China exhibits noticeable spatial dependence, but in distinct ways. The working patterns in the US are influenced by those of their neighbors (spatial clustering of observed outcome) whereas the unobserved characteristics of Chinese workers are correlated with their neighbors that can explain the difference in work culture (spatial clustering of unobservable). Conversely, the business day work intensity of GitHub users in Germany and India does not demonstrate significant spatial dependence.

Another noteworthy finding pertains to the coefficients of *Freelancer* across the four countries. The OLS and M2 estimations of *Freelancer* in China, as well as the M1 estimation of *Freelancer* in India, are positive and significant, whereas the coefficients in the remaining countries are negative. This indicates that unemployed GitHub users in China and India tend to be more actively engaged in open collaboration communities compared to their employed counterparts. One plausible explanation for this discrepancy is that, in developing Asian countries like China and India, open collaboration culture may be less prevalent and popular. Commercial institutions in these regions are more inclined to consider their products and

codes as private properties, thus limiting their distribution on public knowledge-sharing platforms. Consequently, individuals employed by these institutions participate less actively in open collaboration communities.

When examining the regression analysis of *Weekend_Work_Intensity*, $\rho^s$ is significant only in the United States, while $\lambda^s$ is significant solely in India. This implies that the weekend work intensity of GitHub users in the US and India displays geographic spatial interdependence, but also in distinct manners. Among the four countries analyzed, only the coefficients of *Freelancer* in China demonstrate significance. This indicates that unemployed GitHub users in China exhibit a significantly higher level of engagement in open collaboration compared to the employed. This finding can be partially attributed to the relatively low prevalence of an openness culture in China. Additionally, it may suggest differences in the popular usage patterns of the GitHub platform in comparison to the other countries.

Finally, for *Difference_Work_Intensity*, both $\rho^s$ and $\lambda^s$ are significant in China, and $\rho^s$ is significant in Germany, indicating that people's variation of work intensity between business days and weekends are geographically associated in the two countries. When examining the three dependent variables as indicators of distinct aspects of GitHub users' work patterns, it is noteworthy that while the variables associated with geographical spatial effects do not consistently exhibit significance like the digital spatial effect indicator, we can still affirm that the work patterns of GitHub users in all four countries are influenced to varying degrees by their geographical distribution. Despite the diverse ways in which this influence manifests, it is evident that geographical factors play a role in shaping the work patterns of GitHub users across the four countries, thereby supporting the proposed social

superposition state assumption.