Timezone and Time-of-Day Variance in GitHub Teams: An Empirical Method and Study

Prem Devanbu,*† Pallavi Kudigrama,† Cindy Rubio-González,† Bogdan Vasilescu‡
†University of California, Davis, USA ‡Carnegie Mellon University, USA
{ptdevanbu, pkudigrama, crubio}@ucdavis.edu, vasilescu@cmu.edu

ABSTRACT

Open source projects based in ecosystems like GitHub seamlessly allow distributed software development. Contributors to some GitHub projects may originate from many different timezones; in others they may all reside in just one timezone. How might this timezone dispersion (or concentration) affect the diurnal distribution of work activity in these projects? In commercial projects, there has been a desire to use top-down management and work allocation to exploit timezone dispersion of project teams, to engender a more round-the-clock work cycle. We focus on GitHub, and explore the relationship between timezone dispersion and work activity dispersion. We find that while time-of-day work activity dispersion is indeed associated strongly with timezone dispersion, it is equally (if not more strongly) affected by project team size.

CCS CONCEPTS

• **Social and professional topics → Geo characteristics**;

KEYWORDS

GitHub, Timezones, Circular Statistics

ACM Reference format:
Prem Devanbu, Pallavi Kudigrama, Cindy Rubio-González, Bogdan Vasilescu. 2017. Timezone and Time-of-Day Variance in GitHub Teams: An Empirical Method and Study. In *Proceedings of 3rd International Workshop on Software Analytics, Paderborn, Germany, September 4, 2017 (SWAN'17),* 4 pages.
https://doi.org/10.1145/3121257.3121261

1 INTRODUCTION

Coding platforms like GitHub facilitate distributed development. With cloud-based version control, branching, merging, and even build and test capability, it is possible for widely dispersed team members to actively contribute to software projects. Geographic dispersion has traditionally been feared to be an impediment to productivity and quality; but thanks to such platforms as GitHub and the widespread use of video conferencing facilities, Colazo argues [3] that geographic dispersion should be superseded by the concept of *virtual proximity*, where the geographic separation has been largely overcome by modern communication technologies.

Others have noted that geographical dispersion associates with temporal dispersion [3], which can cause coordination and communication overhead, with adverse impacts on interval [9] and software quality [2]. However, when teams are dispersed over many timezones, they can "follow the sun", so that developers take advantage of time differences to sustain work around the clock, while also being able to enjoy normal wake-sleep cycles. "Follow the sun" is considered a significant advantage [4] in *commercial* projects; this might for example allow for better time-to-market, and perhaps more rapid response to bug reports, customer requests, etc.

We focus on GitHub open-source software (OSS), rather than on purely commercial projects. While GitHub allows OSS projects to be geographically and temporally dispersed, OSS projects are also usually more free-wheeling than commercial projects, with reduced intensities of top-down management and control. Given this, how does timezone dispersion of developers affect time-of-day (ToD) of contributions? Does a more geographically distributed contributor pool lead to more around-the-clock development?

The quantitative study of temporal phenomena distributed along diurnal cycles must consider the fact that hours are cyclical. Thus +0800 and +1000 timezones are exactly as proximate as +1400 and -1100, although numerically it may not appear so. Thus, the average of +0800 and +1000 is 0900; the average of the latter pair is not Greenwich Mean Timezone, but rather at the International Date line. One has to resort to special types of statistics to calculate distributional moments such as mean, variance, concentration, skew, etc. To our knowledge, we are the first to study GitHub team work cycles using circular statistics. Our contributions are:

- We gather the timezones and contribution times (UTC) from several large GitHub projects (Section 2).
- We introduce the use of *circular statistics* to study the dispersion of timezones and work times (Section 2).
- We evaluate how timezone dispersion affects work hour dispersion (Section 3).

We discus the threats to validity in Section 4, related work in Section 5, and conclude in Section 6.

2 METHODS

We collected and statistically analyzed data from a sample of open-source projects on GitHub, as described below.

2.1 Data Gathering

Git commit logs provide both a timestamp (UTC as per GitHub) and a timezone (from the committer's computer configuration) [8]. We mined data from a sample of 223 large GitHub projects, spread

*Authors are listed alphabetically.

SWAN'17, September 4, 2017, Paderborn, Germany

ACM ISBN 978-1-4503-5157-7/17/09…$15.00
https://doi.org/10.1145/3121257.3121261

Table 1: Summary statistics (223 projects/rows).

Statistic	Mean	St. Dev.	Min	Median	Max
ToD_variance	0.52	0.14	0.24	0.52	0.83
Timezone_variance	0.22	0.18	0.01	0.19	0.83
Commits	18, 452.04	14, 762.31	1, 770	13, 443	98, 891
Contributors	333.31	553.82	10	157	4, 241

across 6 popular programming languages: C, Java, Javascript, PHP, Python, and Ruby, selected as follows. We started with a list of non-fork and non-mirror projects with at least 5,000 commits each, as recorded in GHTorrent [10], from which we arbitrarily sampled. We then used Perceval[1] to extract commit timestamps and timezone data. We extracted the ToD and the timezone from the commit logs. Timezones are specified as a time difference from GMT, in integer hours; for the sake of simplicity, we round non-integer time zones, such as GMT+5:30, used in India, to their floor integer [8]. We built *project ToD profiles*, reflecting the distribution of commits at different UTC times-of-day, across the 24 hours. We also built *project timezone profiles*, similarly, using the distribution of commits across 27 timezones[2] (-1200 to +1400).

We also observed that timezone metadata is lost in git logs for projects that started in SVN and later migrated to git[3] (and GitHub). We conservatively removed these projects from further consideration. We further performed identity matching (dealiasing) to link different aliases used by a contributor using heuristics from prior work [7], such that we could accurately estimate the total number of contributors to each project. Finally, we performed outlier detection and further excluded 13 projects identified as outliers on #Commits, #Contributors, or ToD and Timezone dispersion (details on measures below). The final sample (223 projects; Table 1) reflects all these filters.

2.2 Time of Day and Timezone Profiles

We emphasize that both time-related attributes we consider (timezone and commit time-of-day, ToD) have *circular distributions* (*e.g.*, ToD is on a 24-hour clock). This becomes clear when looking at the circular histogram in Fig. 1; we show the distribution of commit ToD from which the commits originated, in two projects (conventional work hours, 9am-5pm UTC, are shown in dark blue). Circular random variables follow modular arithmetic, so the calculation of moments such as means, variances etc, is done differently, essentially by integrating around a polar (angular) density function.

Our primary interest here is the *dispersion* of the timezone origin of contributions, and the corresponding effect (if any) on the ToD dispersion of work (commits). To measure the timezone dispersion, we use the second circular moment, also known as *circular variance*, of the timezone data (#Commits per timezone, over 27 timezones, -1200 to +1400). Variance is a well known measure of dispersion of a distribution. Circular variance works as expected for hours: thus a project where all contributions come equally from +0800 and +0900 timezones will have the same circular variance as those that come equally from +1400 and -1200 time zones, which would naturally be smaller than the circular variance of a project whose contributions all arrive equally split between the +0300 and +0600 timezones. We compute the ToD dispersion (circular variance) similarly, on the ToD data (#Commits/ToD, over 24 hours). We use R's `Directional` and `circular` packages.

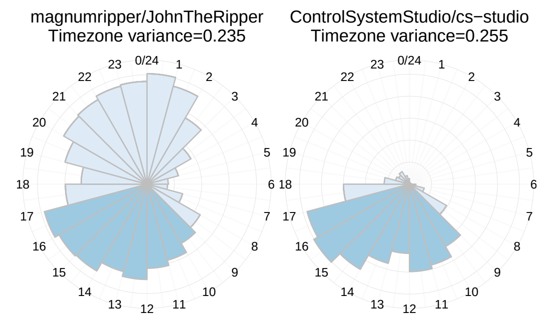

Figure 1: Commits by UTC time of day. Both projects are comparably spread across timezones (timezone variance equals 0.235 and 0.255, respectively; range in our dataset is 0.007 to 0.835), yet have very different commit time of day distributions. UTC business hours 9am-5pm highlighted.

Summary statistics are shown in Table 1. We consider 223 projects with 10 to 4,241 contributors, and 1,770[4] to 98,891 commits. The top two rows show summary statistics for the *circular variance* values for each project. Some projects have quite low circular variance of timezone (0.01 radians^2, indicating that virtually all commits originate in a single time zone; others are quite high, at 0.83 radians^2). Note that timezone and commit ToD dispersion are not necessarily correlated: the two projects in Fig. 1 are comparably dispersed across timezones (circular timezone variance equals 0.235 and 0.255, respectively), yet have very different commit ToD distributions (circular ToD variance equals 0.346 and 0.826, respectively).

2.3 Statistical Analysis

We modeled the variability in timezone dispersion per project, as dependent on UTC ToD dispersion and number of contributors, while controlling for project size (measured as total number of commits; larger projects are likely to behave differently) and programming language (we use GitHub's repository label; different communities, proxied by the language, may have different culture / contributors with different demographics and work styles). Collinearity among predictors was assessed using the variance inflation factor (VIF) and found not to be significant (all below 2).

Then we fitted a linear mixed-effects model with a random-effects term for language. This allows us to capture language-to-language variability in the response to control for potential cultural differences, rather than assessing the effect of specific languages. All other variables were modeled as fixed effects. We used multiple linear mixed-effects models (`lmer` and `lmer.test` in R). Modeling assumptions hold: the QQ-plot did not show significant deviation from a normal distribution; residuals between the observed and model fitted values appeared randomly distributed across the range.

[1] https://github.com/grimoirelab/perceval
[2] https://en.wikipedia.org/wiki/Time_zone
[3] These projects tend to maintain a `git-svn-id` for their pre-git commits in the git logs, which can be used to detect migration.

[4] We note an inconsistency relative to GHTorrent, which reports more than 5,000 commits for all our projects; this could be due to changes to the repositories on GitHub subsequent to the GHTorrent mirroring.

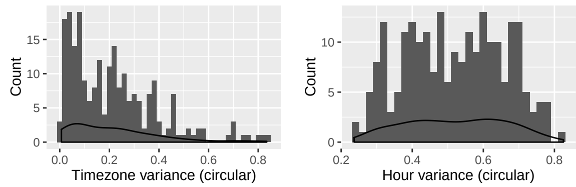

Figure 2: The ToD circular variance between projects is greater than the timezone circular variance.

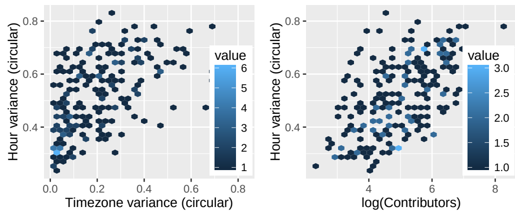

Figure 3: Both timezone circular variance (left) and number of contributors (right) show a positive correlation to ToD circular variance.

Table 2: Time of Day Model: Response is variance(commit ToD). $R^2_m = 0.45$. $R^2_c = 0.48$

	Coeffs (Errors)	Sum Sq.
(Intercept)	$0.727\ (0.107)^{***}$	
Timezone_variance	$0.172\ (0.044)^{***}$	0.98^{***}
log(Commits)	$-0.060\ (0.012)^{***}$	0.04^{***}
log(Contributors)	$0.067\ (0.007)^{***}$	0.94^{***}

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Coefficients are considered important if they were statistically significant ($p < 0.05$). Their effect sizes are obtained from ANOVA analyses. We evaluate our model's fit using a marginal (R^2_m) and a conditional (R^2_c) coefficient of determination for generalized mixed-effects models [12] (`MuMIn` package): R^2_m describes the proportion of variance explained by the fixed effects alone; R^2_c—by the fixed and random effects together.

3 RESULTS AND DISCUSSION

We show the distribution of the timezone and commit ToD circular variance in Fig. 2. On the left, we see that timezone variance is actually not as high. The mode is quite close to zero, suggesting that many projects have very low timezone variance; *i.e.*, their commits mostly originate in a single timezone. Indeed, the median of timezone variance (from Table 1) is just 0.19 radians[2]. On the other hand, the ToD commit variance is almost 3X as high, at 0.53. There are two possible sources for the ToD variance: individual behaviour differences of people within the project (even when working within the same timezone), and natural circadian differences arising from different timezones of origin; additional variance could also arise from greater activity: more activity may influence ToD commit behaviour. Which project aspect is associated more with the often-desired [4] "follow the sun" round-the-clock model, *i.e.*, greater ToD circular dispersion?

Our multiple regression analysis in Table 2 is intended to gauge the relative influence of these factors on commit ToD dispersion. The model suggests that all variables: timezone variance, commit count, and contributor count, have significant effects on the ToD variance. The model has good explanatory power: overall variance explained is 0.48, which is moderately high. The *sum of squares* column on the right is a measure of the variance explained by each variable. Clearly, timezone variance explains about half the variability (sum sq. = 0.98) in the ToD variance; the other (roughly) half, 0.94, comes from the contributors. Commits and language explain a much smaller portion of the variation in the ToD variance. This can also be visually appreciated in the two bivariate plots in Fig. 3. the plot on the right shows a strong, positive relation between log(contributors) and ToD variance, with the points fairly evenly spread around a diagonal line; the plot on the left shows the weaker (but still perceptible) relationship between timezone and ToD variance. Indeed, if one reverses the order of covariates in the regression, we find that timezones explain even less of the variance in hours. The rather weak negative association of commit count with work hour ToD dispersion (weak explanatory power, barely 5% of that of timezones or contributor count) is rather puzzling, and appears inconsistent with earlier studies (recent work [13] finds that "100% of studies reported a positive relationship" between temporal dispersion and productivity). This might be an artifact due to some larger projects with more commits, which may have had most of the commits early in their history originating from Europe or North America.

Our conclusion is that **when it comes ToD dispersion (and thus the prospect of "following the sun") workforce *size* matters as much (or indeed more) than *location*.** Besides showcasing novel methodology to study geographic / temporal dispersion in OSS, this result calls for more research into the coordination challenges in colocated software development: we find that larger same-timezone teams exhibit as much natural diversity in working habits (and potentially as many coordination challenges) as smaller timezone-dispersed teams.

4 THREATS TO VALIDITY

Construct Validity assesses whether the variables we considered accurately model the constructs of our study. Our two main constructs are geographic dispersion (based on timezone metadata from git commit logs) and temporal dispersion (based on timestamp metadata from said logs). We note several validity threats. First, the data could be affected: git commit metadata can be overwritten by users; the timezone metadata may not reflect a developer's actual location if work is carried out through a remote machine; timezone metadata for SVN projects migrated to git may be lost. We addressed the latter by excluding such projects from our dataset. The impact of the former should be assessed through replications on different samples. Second, since both time-related attributes we considered have circular distributions (*e.g.*, 11pm is as far from 1am as 7am is from 9am), dispersion measures for "standard" distributions are inappropriate as aggregation techniques. We addressed this by using specialized circular statistics instead of standard dispersion measures (see Section 2.2). Third, we made a number of simplifying assumptions when collecting and aggregating our data that could have affected our results, *e.g.*, we rounded non-integer timezone offsets, considered all 27 timezones from -1200 to +1400 as equidistant, and did not account for daylight savings zone changes.

Internal Validity is the extent to which the conclusions can be drawn from the conducted measurements. To ensure our results are robust, we systematically dealt with outliers in our data, employed a well-established statistical modeling technique (multiple linear regression), included controls for confounds, and checked if we comply with modeling assumptions. Still, even if robust, our results in no way imply causality, but rather represent a strong statistical correlation between the measured attributes. This could be strengthened through a deep qualitative dive into the specifics of teams in our sample, which is beyond the scope of this paper.

External Validity pertains to generalizability of our findings. The projects we extracted were based on data only from GitHub, which limits the generalizability. To reduce this threat, we ensured diversity in our data by including projects written in six languages, and having different sizes, and different geographic and temporal dispersion profiles.

5 RELATED WORK

The increasing prevalence and significance of distributed software development has motivated numerous studies on the effects of temporal and spatial dispersion.

Temporal dispersion has been viewed as an impediment to coordination and communication [2]. Some papers report negative effects on productivity [9]. However, the "Follow The Sun" idea has been pursued as a way to reduce intervals. Colazo and Fang [4] report that timezone dispersion of teams can both reduce intervals and improve quality, but this relationship is moderated by software complexity. Espinosa et al. [5] report that effects of temporal dispersion on performance are mediated by team interaction effects and timezone overlap. Sivunen et al. [17] argue that the experience of temporal boundaries is not symmetrical to global collaborators and small time differences can sometimes be more challenging than large time differences in global virtual work. This can however be addressed with moderate timeshifting by team members [14].

Geographic dispersion is evidently different from timezone dispersion, and can have different effects. Takhteyev and Hilts [18] found that most GitHub developers seem to be highly clustered around North America and Western and Northern Europe. Gharehyazie and Filkov [6] studied collaboration and its effect on productivity among groups of developers in 26 Apache OSS teams, and tracked the differences in developer behavior when part of a team is remote. Tang et al. [19] identified various strategies used to find windows of time to interact synchronously. Schilling et al. [16] hypothesize that spatial, temporal, and cultural distances are key factors for team integration and project retention, and provide metrics to measure these factors. Bird and Nagappan [1] determined the effect of organizational and geographic distribution on pre- and post-release defects. Work patterns also depend on factors such as work context and concentration levels [11], or renumeration [15].

6 CONCLUSION

GitHub and other platforms are allowing OSS projects to recruit collaborators from around the world, operating from widely dispersed timezones. Timezone dispersion in commercial projects has been considered an advantage, as it can lead to a "follow the sun" work practice, allowing round-the-clock work activities. Does this also hold for OSS projects on GitHub, with the advanced support for distributed development that GitHub provides? We studied 223 projects, with between 10 and 4,241 contributors, and examined the timezone and ToD dispersion of their commits. In most projects, we found fairly modest variance in the timezones from which people originate commits, but much greater variance in the time-of-day at which developers originate commits.

Secondly, we studied the factors that might influence the commit ToD variance. As expected, we found that both the number of committers and the timezone dispersion have positive, strong association with the ToD dispersion. The surprising finding here is that the effects are equally strong. One might have expected that timezone would have the strongest effect. This suggests that variation between people and their working habits have as strong an influence as geographical (timezone) dispersion, on whether the project "follows the sun". Further studies, including a qualitative deep dive into the specific behaviours of projects and individuals, are required to understand the causal factors.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This work was supported in part by NSF grant CNS-1629976.

REFERENCES

[1] C. Bird and N. Nagappan. Who? where? what?: examining distributed development in two large open source projects. In *MSR'12*, pages 237–246.
[2] M. Cataldo. Sources of errors in distributed development projects: implications for collaborative tools. In *CSCW'10*, pages 281–290.
[3] J. Colazo. Structural changes associated with the temporal dispersion of teams: Evidence from open source software projects. In *HICSS'14*, pages 300–309.
[4] J. A. Colazo and Y. Fang. Following the sun: Temporal dispersion and performance in open source software project teams. *JAIS'10*.
[5] J. A. Espinosa, N. Nan, and E. Carmel. Temporal distance, communication patterns, and task performance in teams. *JMIS'15*, pages 151–191.
[6] M. Gharehyazie and V. Filkov. Tracing distributed collaborative development in apache software projects. *ESE'16*, pages 1–36.
[7] M. Gharehyazie, D. Posnett, B. Vasilescu, and V. Filkov. Developer initiation and social interactions in OSS: A case study of the Apache Software Foundation. *Empirical Software Engineering*, 20(5):1318–1353, 2015.
[8] J. M. Gonzalez-Barahona, G. Robles, and D. Izquierdo-Cortazar. Determining the geographical distribution of a community by means of a time-zone analysis. In *OpenSym'16*.
[9] A. Gopal, J. A. Espinosa, S. Gosain, and D. P. Darcy. Coordination and performance in global software service delivery. *IEEE-TEM'11*, pages 772–785, 2011.
[10] G. Gousios and D. Spinellis. GHTorrent: Github's data from a firehose. In *MSR'12*, pages 12–21.
[11] G. Mark, S. T. Iqbal, M. Czerwinski, and P. Johns. Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In *CHI'14*, pages 3025–3034.
[12] S. Nakagawa and H. Schielzeth. A general and simple method for obtaining r^2 from generalized linear mixed-effects models. *MEE'13*, pages 133–142.
[13] A. Nguyen-Duc, D. S. Cruzes, and R. Conradi. The impact of global dispersion on coordination, performance and software quality–a systematic literature review. *IST'15*, pages 277–294.
[14] R. Prikladnicki and E. Carmel. Is time-zone proximity an advantage for software development? the case of the brazilian it industry. In *ICSE'13*, pages 973–981.
[15] D. Riehle, P. Riemer, C. Kolassa, and M. Schmidt. Paid vs. volunteer work in open source. In *HICSS'14*, pages 3286–3295.
[16] A. Schilling, S. Laumer, and T. Weitzel. Together but apart: how spatial, temporal and cultural distances affect floss developers' project retention. In *CPR'13*.
[17] A. Sivunen, N. Nurmi, and J. Koroma. When a one-hour time difference is too much: Temporal boundaries in global virtual work. In *HICSS'16*, pages 511–520.
[18] Y. Takhteyev and A. Hilts. Investigating the geography of open source software through github, 2010.
[19] J. C. Tang, C. Zhao, X. Cao, and K. Inkpen. Your time zone or mine?: a study of globally time zone-shifted collaboration. In *CSCW'11*, pages 235–244.