Influence on GitHub: Individual Limits and Organization Advantages

Ian Dennis Miller
University of Toronto
Psychology Department
100 St. George St. 4th Floor
Toronto ON, M5S 3G3
Canada

Abstract

gh-impact is a new measure of influence on GitHub. In this paper, we introduce *gh-impact* and use it to examine differences between Individual and Organizational GitHub accounts. Among our key findings, we find *gh-impact* increases as the size of organizations increase, suggesting that individuals may have a comparative disadvantage. We also find evidence for a ceiling effect of the impact of Individuals. While some individuals manage extremely successful projects, it is rare to find Individuals who manage multiple projects of a similar caliber. Organizations are not inherently resource-bound and can sustain many projects in parallel, leading to greater overall work impact and a correspondingly higher *gh-impact* score. *gh-impact* scores can be explored online at **http://www.gh-impact.com**

Introducing *gh-impact*

gh-impact is a measure of influence on GitHub. *gh-impact* is based upon the stars a project receives: an account has a *gh-impact* score of n if they have n projects with n stars. Our *gh-impact* formulation is quite similar to that of the academic H-Index (Hirsch 2005), but we utilize Project Stars as our measure of engagement instead of article citations. Higher *gh-impact* scores correspond to accounts that have many well-used projects. *gh-impact* can provide a rough estimate of a GitHub account's overall productivity and impact.

Methods

This work makes extensive use of the GHTorrent archives provided by Gousios (2013). We included all users ($n = 13,203,696$), projects ($n = 34,672,644$), organization memberships ($n = 367,319$), follows ($n = 11,616,754$), and stars ($n = 49,243,032$) that were present in the July 17, 2016 data dump. Data were imported into a Postgres database (Stonebraker and Kemnitz 1991) for preprocessing before loading it into R (RDevelopment Core Team 2008). *gh-impact* itself is expressed as a series of SQL views.

Results

Our *gh-impact* calculations ultimately yielded $n = 1,064,714$ accounts with a *gh-impact* score above 0. Of these accounts, $n = 918,061$ are individuals and $n = 146,652$ are organizations. Organizations ($mean = 1.86$) tend to have higher *gh-impact* scores than Individuals ($mean = 1.61$). The cumulative distribution of *gh-impact* rapidly tops out; the $10,266$ accounts with scores above 8 are in the 99th percentile.

Popularity Penalty for Individuals

To explain the discrepancy in *gh-impact* between Individuals and Organizations, we controlled for quantity of projects, total stars received, total followers, and stars received by the account's most popular project. We found that Individual accounts with a very popular project, as a proportion of their total stars, had lower *gh-impact* scores.

In Figure 1, we used a Monte Carlo Bootstrap method to estimate size of the penalty effect for Individuals. To see this effect a different way, we plotted *gh-impact* against each account's most popular project in Figure 2. Although Organizations continue to gain *gh-impact* as their projects become more popular, Individuals present a ceiling effect that no amount of popularity can overcome.

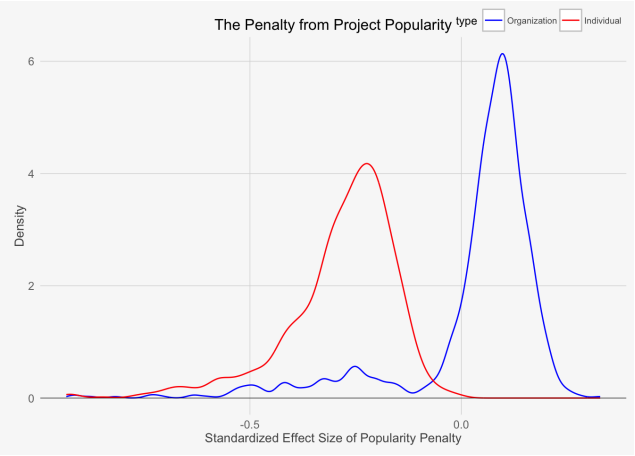

Figure 1: The penalty effect for Individuals is significantly below 0. For Organizations, the penalty effect is not significant.

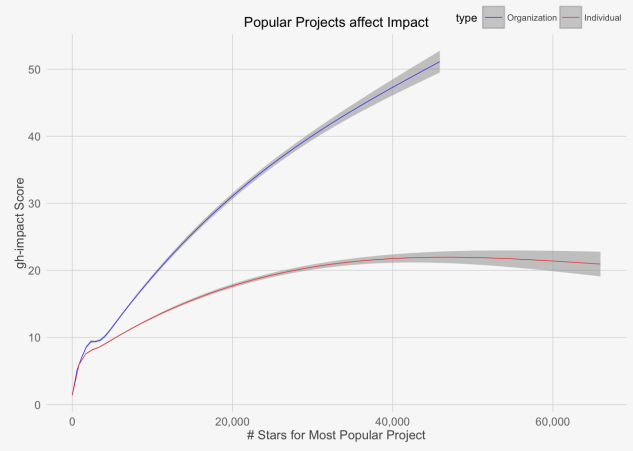

Figure 2: Individuals *gh-impact* appears to stop increasing around $ghi = 20$. Organizations are not bounded on *gh-impact*.

Organization Impact and Organization Size

Organizations can help coordinate work, but scaling problems can also lead to disorganization and work loss. If the GitHub platform successfully helps Organizations cope with growth, then we would expect to see increases in *gh-impact* as organization size grows. The Pearson's correlation of *gh-impact* and membership is significant, $r = 0.51, (p < 0.01)$. This is a large effect (Cohen 1992) that can be seen in Figure 3. The discontinuity observed in the middle of the distribution can be attributed to several small organizations with unusually high *gh-impact scores*. Facebook ($ghi = 147$), Mozilla ($ghi = 95$), and Twitter ($ghi = 88$) have relatively higher *gh-impact* scores despite having between 250-400 organization members.

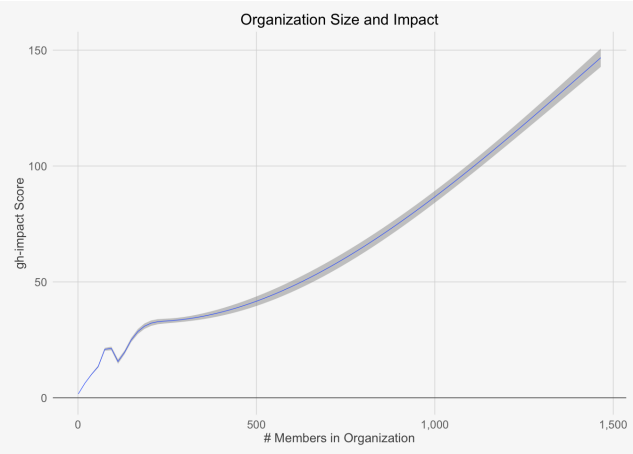

Figure 3: As organization size grows, impact also grows.

Limited Social Reciprocity on GitHub

It is commonly believed in academia that some authors use citations strategically, leading to biased *H-Index* scores. Could the same be true among GitHub users? To test whether users starred projects for purely social reasons, we looked for reciprocity among social behavior on GitHub.

If giving and receiving were reciprocal, we would expect to find a correlation of 1. Instead, the Pearson's correlation between stars given and stars received is $r = 0.05, (p < 0.01)$ for Individuals and $r = 0.03, (p < 0.01)$ for Organizations. Since both effects are "smaller than small" (Cohen 1992), there is evidence that starring reciprocity does not exist. In Figure 4, there may be some evidence of reciprocity among Individuals with very little activity because they distribute more stars than they receive. Among Organizations, there is no social reciprocity at any point. Thus, it does not appear that projects are starred for the purpose of social reciprocity.

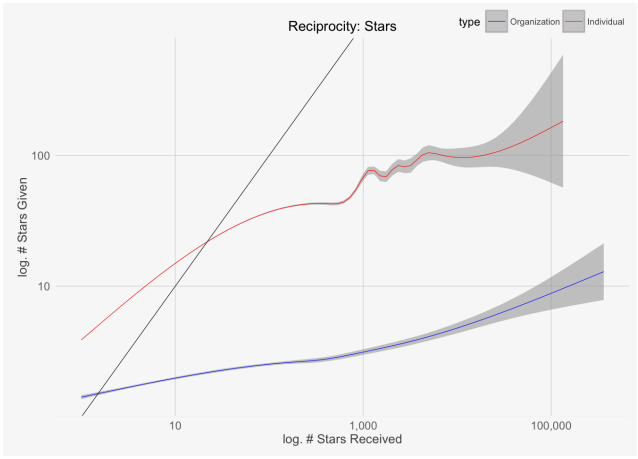

Figure 4: A solid black line with $slope = 1$ represents the perfect reciprocity scenario: each star received would be matched by a star given. Accounts below this line receive more stars than they give.

Brief Discussion

We wonder whether stars serve a functional role, rather than a social role, which could explain why stars are a good substitute for academic citations in our formulation of *gh-impact*. To explain the "Popularity Penalty," we wonder whether a single demanding project can inhibit the creation of new projects, thereby leading to an overall lower *gh-impact* score. [1] Organizations do not have the same attentional resource constraints.

Conclusion

gh-impact has many interesting properties that make it useful for characterizing GitHub accounts and for investigating collaboration dynamics. Findings will be posted to **http://www.gh-impact.com** as this work nears publication in an archival journal.

[1] For example, Linus Torvalds has a *gh-impact* score of 2, but one of his projects is extremely popular.

References

Cohen, J. 1992. A power primer. *Psychological bulletin* 112(1):155.

Gousios, G. 2013. The GHTorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, 233–236. Piscataway, NJ, USA: IEEE Press.

Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America* 102(46):16569–16572.

RDevelopment Core Team. 2008. R: A language and environment for statistical computing. *R Foundation Statistical Computing*.

Stonebraker, M., and Kemnitz, G. 1991. The POSTGRES next generation database management system. *Communications of the ACM* 34(10):78–92.