Effects of Gender on Contribution Evaluation on Github

Pascal Brokmeier (B.Sc.)
Universität zu Köln, Germany - Email: pbrokmei@smail.uni-koeln.de

Abstract

!!!TODO!!!

1. Introduction

Software development has adapted the concept of social coding and pull based software development through platforms like Github which provides a platform for some of the biggest Open Source Software (OSS) projects existing. Projects like Ruby, NodeJS, Bootstrap, Angular or the Java Spring Framework are publicly hosted with some of them having thousands of followers and contributors [1]. Open Source software development has been described as meritocracies [2], however recent research has identified social factors to influence decisions of project managers regarding the acceptance of contributions by others [3]. Inevitably, a social coding environment such as Github is accompanied with social interaction that influences project progress.

An obvious factor in social interaction is gender. Sociology research has shown that women are being treated unequal in professional environments !!!GOOD CITE NEEDED!!! and more specifically OSS projects have been shown to exhibit sexist behavior. About 1.5% of the total number of members in communities of *'free/libre/open source software (F/LOSS)'* were determined to be female compared to 28% in proprietary software as determined in a 2005 report by the University of Cambridge [4]. More current research shows a percentage of about 9% female users on Github [5]. Since Tsay et al. found that project managers use social cues to evaluate contributions and Vasilescu et al. found that almost half of the project members are aware of other users gender, the effect of the perceived gender on this contribution process can be of interest in the ongoing debate of gender inequality.

These factors, the underrepresentation of women on Github, the observed sexist behavior within OSS communities as well as the social influences on decisions that were believed to be purely lead by meritocratic reasoning raise the following question:

How does the perceived gender affect the evaluation of contributions by members in a public social coding environment such as Github?

1.1. Method and expected results

In order to answer the research question, three subtasks can be identified that need to be completed in order for the question to be answered adequately

- How can the perceived gender for each user be determined? What is the perceived gender for each user?
- When is a contribution considered to be accepted? When is it considered to be rejected?
- Is a correlation existing between gender of the contributor and the contribution evaluation result? Can causal identification be achieved?

The first two tasks need to be resolved using proper technical analysis of the data and using data such as followers as proxies for social standing and network embeddedness. The last task is reliant on the results of the first two as well as the methods available to us to achieve causal identification such as Relational Covariate Adjustment (RCA). While a simple correlation would already be of interest, discovering a causal effect would be more satisfying.

To analyze the data, the GHTorrent !!CITE!! dataset is used, which allows the execution of queries against a huge dataset of all repositories, users, pull-requests, comments and issues on Github since 2013. The dataset included over 14 million users and 13 million pull requests in November 2016 although the number of relevant users is expected to be much lower since the distribution of active users is a long tail distribution with very few active users and many inactive or abandoned accounts. Nonetheless, the scale of the dataset allows for thorough data preprocessing without loosing too many potential entries for the analysis afterwards.

To determine the perceived gender, the genderComputer by Vasilescu et al. will be used. This algorithm uses several heuristics such as the origin/country of the user as well as common name patterns and a big name-gender dictionary to infer the gender from a given user. It has a reported success rate of about 32% [5].

To rate the evaluation of a contribution, we consider a merged pull request (PR) to be an accepted contribution and a closed but not merged PR to be a rejected contribution. Since PR are hard to determine as either merged or not merged however, this information will have to be resolved by investigating the repositories commit history and comparing the commit IDs of the PR with the commit IDs of the repository itself. If the PR contains commit IDs that are also present in the repository itself, the PR can be considered

merged. If they are not included, the PR can be considered rejected [5]..

Tsay et al.

1.2. Data Preparation

1) downloading data from GHTorrent
2) clean all unneeded fields from documents to reduce file size
   - remove forks [7]
   -
   - Repos before removing forks: 30713080
   - Repos after removing forks:
   - remove repos without at least 1 fork
   - repos before:
   - repos after:
3) determine gender of users
4) remove inactive users
5) filter through repositories, deleting inactive or small ones [7]
6) (select biggest repositories)

```
//remove all forked repositories
db.repos.remove({fork: true}, false)
//remove all repositories that have no forks (and
  therefore no chance for pull requests)
db.repos.remove(\{forks: \{: 1\}\}, false)
```

2. Data Processing

1) iterate over pull requests
2) for each: determine if accepted / rejected

1)

3. Conclusion

References

[1] G. Gousios, "The ghtorrent dataset and tool suite," in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 233–236. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487085.2487132

[2] W. Scacchi, "Free/open source software development: Recent research results and emerging opportunities," in *The 6th Joint Meeting on European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering: Companion Papers*, ser. ESEC-FSE companion '07. New York, NY, USA: ACM, 2007, pp. 459–468. [Online]. Available: http://doi.acm.org/10.1145/1295014.1295019

[3] J. Tsay, L. Dabbish, and J. Herbsleb, "Influence of social and technical factors for evaluating contribution in github," in *Proceedings of the 36th International Conference on Software Engineering*, ser. ICSE 2014. New York, NY, USA: ACM, 2014, pp. 356–366. [Online]. Available: http://doi.acm.org/10.1145/2568225.2568315

[4] B. Krieger and J. Leach, "Free/libre and open source software: Policy support - gender: Integrated report of findings," University of Cambridge, Tech. Rep., 2006. [Online]. Available: http://flosspols.merit.unu.edu/deliverables/FLOSSPOLS-D16-Gender_Integrated_Report_of_Findings.pdf

[5] B. Vasilescu, D. Posnett, B. Ray, M. G. van den Brand, A. Serebrenik, P. Devanbu, and V. Filkov, "Gender and tenure diversity in github teams," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: ACM, 2015, pp. 3789–3798. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702549

[6] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study of stackoverflow," in *Social Informatics (SocialInformatics), 2012 International Conference on*, Dec 2012, pp. 332–338.

[7] G. Gousios, M. Pinzger, and A. v. Deursen, "An exploratory study of the pull-based software development model," in *Proceedings of the 36th International Conference on Software Engineering*, ser. ICSE 2014. New York, NY, USA: ACM, 2014, pp. 345–355. [Online]. Available: http://doi.acm.org/10.1145/2568225.2568260