EasyChair Preprint

№ 5981

Trouble in Programmer's Paradise:
Gender-Biases in Interacting on Stack Overflow

Siân Brooke

July 1, 2021

Trouble in Programmer's Paradise: Gender-Biases in Interacting on Stack Overflow

Siân Brooke
s.j.brooke@lse.ac.uk
Department of Methodology, London School of Economics and Political Science

Often called a 'programmer's paradise', Stack Overflow (SO) is a question-and-answer based website focused on writing computer code. It is the largest online community of programming knowledge, boasting over 14 million registered users and more than 50 million monthly visitors (Stack Overflow, 2020). Its convenience and accessibility are typical of the 'open source' nature of coding culture, so often proclaimed as an open meritocracy (Wajcman, 2009). SO covers a range of topics and levels of technical familiarity, from complete beginners to seasoned programmers dealing with questions of algorithmic complexity. With low material barriers to entry, anyone in theory can join this forum, allowing enthusiasts and professionals to meet with a shared passion for 'open' technology. SO (2020) is a nexus of computer-mediated communication and programming knowledge where participation can have real consequences for employment, drawing people from all walks of life to hone their coding skills. The image of an open, anonymous forum is fitting with early, utopian visions of the internet (Wajcman, 2009). But although it may be paradise to some, SO is plagued by the misogyny that defines technology culture (Ford et al., 2017). Examining gendered expression in usernames, this study investigates if and how gender mediates and dictates the recognition of technical knowledge on SO. I ask; *Is Stack Overflow organised by gender?*

In looking to existing scholarship on SO, the role of gender in reputation and activity has been analyzed on an individual level (Ford et al., 2016; Vasilescu et al., 2012), however the effect of gender on interaction dynamics has yet to be studied. Scholars have employed social network analysis to examine geographies of participation on SO (Stephany et al., 2020), but not gender. Using interviews, Ford et al. (2016) determined that the main barrier to women's participation is feeling that they lack the adequate technical qualifications to contribute. In a follow-up study, Ford et al. (2017) also found that women are more likely to participate in a conversation on SO if they see other women already taking part, or even if women are just a visible presence. They referred to the positive influence of similar others as *peer parity* (Ford et al., 2017). This study empirically tests the notion of gendered peer parity, extending analysis to include a non-binary classification of gender. When environments like SO are known as 'boys clubs' this is part of a vicious cycle, which excludes anyone not performing this masculine identity (Ford et al., 2016, p. 6). In identifying gender-bias as it pervades this vital skill-building environment, this paper works towards making technical knowledge open to all.

Data and Methods

The SO data was retrieved from the Stack Exchange Data Dump, a quarterly upload of all the site's content hosted on the Internet Archive. The dataset spans from the founding of the site in 2008 to the most recent quarterly 'dump' when the data was accessed in November 2019. The sampling frame included America, Canada, and the UK. The common cultural heritage of these countries means that they broadly share conceptions of gender, allowing for a measure of generalization. Computational work traditionally relies on a binary understanding of gender, updating the language of *male/female* for a more critical framing in *masculine/feminine* (Risman, 2004). Expanding how gender is operationalised is key to interdisciplinary research that combines a critical paradigm with computational methods. Rather than reproducing the binary this study focuses on 'gender salience', or intensity of gendered self-expression, spanning five-categories: *masculine* ($M = 99,933$), *mostly masculine* ($MM = 45,169$), *anonymous* ($A = 3,864$), *mostly feminine* ($MF = 3,412$), *feminine* ($F = 19,275$). This framing permits the analysis to capture if gender-biases in technical knowledge sharing are intensified with extreme poles (masculine/feminine) of gender representation. In using inclusive gender identification methods, researchers can rely on self-disclosed information, unfortunately SO profiles do not provide this data. Instead I focused on usernames, a user-centred alternative to gathering information on a particular gender. Gender is inferred as masculine or feminine only if a SO user's display name is a 'real name' (given name, such as Jane Smith). Work into gender and SO most commonly uses *genderComputer*, a tool developed by Vasilescu, Capiluppi, and Serebrenik (2012). genderComputer is written in Python, built on a database of first names from 73 countries, as well as gender.c, an open source C program for name-based gender inference. This library is used by a wide range of studies into gender and SO (Ford et al., 2016). Inferring a user's gender from their name and location, genderComputer is specifically built from Drupal, WordPress, and SO. The framing of gender in this study emphasises how automated inference results in a representation of gender, rather than the discovery of a ground truth. The application of non-binary gender inference in this study is a novel contribution to computational work.

Results and Discussion

There were 171,653 users connected in the SO graph: M: 58%, MM: 26%, A: 2%, MF: 2%, and F: 11%. Connections (degrees) are based on questions, answers, and comments for the between users. Unsurprisingly, masculine users have the highest average degree, with the frequency of connections decreasing with the salience of masculinity. Feminine users are the least connected, with the least average number of degrees. This could be because masculine users who use their real names on platforms (which were more easily picked up by the gender classifier) are more likely to be invested in participating in the culture and less likely to face gendered hostility. Masculine and mostly masculine usernames are more likely overall to receive a comment or answer on their post in comparison to other users. Supporting the hypothesis that gender salience disadvantages feminine users as they receive the least responses.

To further understand gendered patterns, we can compare the probability of following a randomly selected edge and arriving at a particular gender. These probabilities are represented by $p(g)$. Though the network consists of considerably more masculine than feminine users, the average masculine in-degree is 16.33, larger than the average feminine in-degree of 11.67, resulting in $p(M) > p(F)$. We can also examine the gendered make-up of a user's neighbourhood, that is, whether they are more likely to be connected to a user of the same gender grouping. Here I am computing the conditional probability $p(g'|g)$ that a random neighbour of individuals with gender g has gender g'. For neighbours of masculine users, $p(F|M) = 0.08$ and $p(M|M) = 0.60$. For feminine users $p(M|F) = 0.55$ and $p(F|F) = 0.14$. In all cases, the random neighbour is more likely to be masculine. However, a random neighbour for a feminine user is almost twice as likely to be another feminine user than for masculine users. This shows evidence of *peer parity* in gender categorisations, even with a majority male population. This illustrates that there is a preference for interacting within gendered boundaries, and the effect is reasonably strong. In total 24% of interactions were reciprocated across the network. I found evidence for gendered homophily as for each gender categorisation, the highest level of reciprocity is from the same gender. The margins for this majority are large, up to three times the reciprocity for any other group. The highest levels of in-group reciprocity are for users identified as mostly feminine and anonymous, whilst the comparatively lower reciprocity of feminine is users is likely due to lower levels of participation in general. The high levels of reciprocity for anonymous and mostly feminine users may reflect a functional use of pseudonyms, where varying degrees of anonymity are used as a resource for legitimate participation. Supporting Ford et al (2016) with empirical evidence, I peer parity is a significant feature of interaction on SO, and this is clear even when gender is operationalised beyond a binary variable.

Conclusion

This study concludes finds that Stack Overflow is fundamentally organised by gender. Even when it does not inform the sampling criteria, gender defines the patterns of participation on SO as a manifestation of wider technology culture. The result holds when gender is operationalised beyond a binary, even with the inclusion of the anonymous categorisation. Such empirical evidence for gendered patterns of interaction challenges meritocratic assumption that level of participation is the sole determinant of reception and sharing on SO. This study shows that merely signalling gender in usernames is sufficient to determine patterns of interactions. As shown by Ford et al. (2016), women do not participate on SO because they fear condescension and antagonism from the majority masculine population. In examining participation, scholars need to ask what barriers are erected by visible gendered boundaries and a homogenous masculine identity. For activists, the path forward lies in publicising the current disparity and simply advocating for openness.

References

Ford, D., Harkins, A., & Parnin, C. (2017). Someone like me: How does peer parity influence participation of women on stack overflow? *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC*, *2017-Octob*, 239–243. https://doi.org/10.1109/VLHCC.2017.8103473

Ford, D., Smith, J., Guo, P. J., & Parnin, C. (2016). Paradise Unplugged: Identifying Barriers for Female Participation on Stack Overflow. *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering - FSE 2016*, 846–857. https://doi.org/10.1145/2950290.2950331

Stack Overflow. (2020). *Stack Overflow Developer Survey 2020*. https://insights.stackoverflow.com/survey/2020#community-developers-perspectives-by-gender

Stephany, F., Braesemann, F., & Graham, M. (2020). Coding together–coding alone: the role of trust in collaborative programming. *Information Communication and Society*. https://doi.org/10.1080/1369118X.2020.1749699

Vasilescu, B., Capiluppi, A., & Serebrenik, A. (2012). Gender, Representation and Online Participation: A Quantitative Study. *The British Computer Society*. https://doi.org/10.1093/iwcomp/xxxxxx

Wajcman, J. (2009). Feminist theories of technology. *Cambridge Journal of Economics*, *34*(1), 143–152.