Anna May and Johannes Wachs and Anikó Hannák

Gender differences in participation and reward on Stack Overflow

Article (Published)
(Refereed)

Empirical Software Engineering
https://doi.org/10.1007/s10664-019-09685-x

Gender differences in participation and reward on Stack Overflow

Anna May[1] · **Johannes Wachs**[2] · **Anikó Hannák**[3,4]

Published online: 08 February 2019

Abstract
Programming is a valuable skill in the labor market, making the underrepresentation of women in computing an increasingly important issue. Online question and answer platforms serve a dual purpose in this field: they form a body of knowledge useful as a reference and learning tool, and they provide opportunities for individuals to demonstrate credible, verifiable expertise. Issues, such as male-oriented site design or overrepresentation of men among the site's elite may therefore compound the issue of women's underrepresentation in IT. In this paper we audit the differences in behavior and outcomes between men and women on Stack Overflow, the most popular of these Q&A sites. We observe significant differences in how men and women participate in the platform and how successful they are. For example, the average woman has roughly half of the reputation points, the primary measure of success on the site, of the average man. Using an Oaxaca-Blinder decomposition, an econometric technique commonly applied to analyze differences in wages between groups, we find that most of the gap in success between men and women can be explained by differences in their activity on the site and differences in how these activities are rewarded. Specifically, 1) men give more answers than women and 2) are rewarded more for their answers on average, even when controlling for possible confounders such as tenure or buy-in to the site. Women ask more questions and gain more reward per question. We conclude with a hypothetical redesign of the site's scoring system based on these behavioral differences, cutting the reputation gap in half.

Keywords Open-source · Online representation · Gender gap · Empirical measurement

Communicated by: Emerson Murphy-Hill

✉ Anikó Hannák
ancsaaa@gmail.com

Johannes Wachs
johanneswachs@gmail.com

[1] Department of Economics and Business, Central European University, Budapest, Hungary

[2] Department of Network and Data Science, Central European University, Budapest, Hungary

[3] Vienna University of Economics and Business, Welthandelsplatz 1 1, 1020 Vienna, Austria

[4] Complexity Science Hub, Josefstädter Str. 39, 1080 Vienna, Austria

1 Introduction

As coding skills find their way into the basic requirements of many well paying jobs (Glass 2016), the underrepresentation of women in technical fields is becoming an increasingly salient issue (Republic 2014). Recent efforts to reduce this discrepancy are multi-faceted: while communities aimed at teaching girls or women to code focus on issues related to self-confidence and gender stereotypes (Cohoon and Aspray 2006; Ahuja 2002), more IT companies and schools are promoting diversity and fighting discrimination (Clayton and Lynch 2002). Online resources also provide significant and informal opportunities for people who want to learn how to code, from free courses to entire communities for learning, discussing, and collaborating (Glass 2016; Wired 2014; Lerner and Tirole 2002). Two prime examples of the latter are Stack Overflow and GitHub. One would hope that the digital nature of these new "knowledge marketplaces" could democratize knowledge and help to mitigate existing inequalities. Yet, research finds that the opposite has happened. Contribution rates for women in open-source programming communities such as GitHub or Stack Overflow are even lower than their overall presence in the IT labor market.[1] These trends align with findings from studies on other open-source communities and knowledge creation platforms, including Wikipedia and OpenStreetMap (Stephens 2013; Ford 2016; Horvath 2014). These studies formulate a variety of hypotheses to explain this effect, including the impact of gender roles and stereotypes, lower confidence and risk aversion among women, and the asymmetric threat of harrassment.

In this study we explore reasons behind low participation and success rates of women on Stack Overflow, the largest Q&A platform for programming and an important resource in the open source IT world. Over time, Stack Overflow has grown into a large database of knowledge which people use at all stages of learning how to program. Questions vary in difficulty and specificity, and the coverage of topics evolves essentially in sync with coding itself. Beyond being a knowledge base, the site also serves as a social platform, job search site, and recruiting tool - it is an important hub in the IT ecosystem.

We collect a gender-balanced sample of over 20,000 user profiles, and use them to investigate the differences in the participation and success of men and women on the site. We frame our analysis in terms of the following questions. Do men and women have different levels of success on Stack Overflow? If so, is it because of differences in how they participate on the platform? We find that the answer to both questions is yes, and follow up by probing the differences in rewards for different kinds of participation on the platform.

More specifically, we find significant gender gaps in activity: women are more likely to ask questions, while men provide more answers and cast more votes. Votes are positive or negative evaluations of other user's questions and answers. Users gain and lose reputation points, the primary measure of success on the platform, for receiving up and down votes. We also observe that men are significantly more successful on the site, measured by their collection of reputation points. Using the Oaxaca-Blinder decomposition (Oaxaca 1973), a method from economics that to the best of our knowledge has not yet been previously applied to measure gender disparities in online communities, we decompose the outcome differences between men and women in terms of differences in their activity.

[1] 7.6% of the users participating in the 2017 Stack Overflow survey (https://insights.stackoverflow.com/survey/2017) identify themselves as women, while in 2015, 25% of the computing jobs were held by women in the U.S. Ashcraft et al. (2016).

While our models show that question and answer behaviors and other user- and community-level features explain a large portion of the success gap. 11% of the reputation gap remains unexplained.

In the final part of the paper, we explore the consequences of a hypothetical redesign of the site's reward system. The proposed alternative scoring system equalizes the rewards for well-liked questions and answers, a simple and justifiable change which does not penalize any group of users in absolute terms. We find that the median woman is marginally more successful than the median man under this revised success measure, reversing the situation under the current system. However even with our recommendation men are still significantly more successful on average due to their overrepresentation among the top users. The recommendation may alter site dynamics as users will be incentivized to ask more and better questions. Given Stack Overflow's stated aim to build a universal knowledge base, we believe that such a shift in the dynamics is in line with the spirit and goals of the platform.

In general however, our findings suggest that fundamental remedies may be needed in order to encourage women to participate more and in different ways. Given the increasing importance of Stack Overflow and similar sites in both the labor market and knowledge creation, our findings underscore the importance of design decisions and interventions even in well-intentioned and organically grown online communities.

2 Related Work

In this section we outline recent work on gender gaps in IT and on the web. We also survey studies which investigate online platforms to detect and measure inequalities.

Gender gaps and IT In the US a woman earns about 80 cents for every dollar a man earns. Even though the gap has been shrinking since 1960, it is still present at most educational levels and lines of work (Blau and Kahn 2016). The gap is larger within traditionally male-dominated fields such as computing. How do these fields remain male-dominated? Research shows that women are significantly underrepresented in academic fields "believed to require attributes such as brilliance and genius" (Leslie et al. 2015) including computer science. When they do choose to enter these fields, they have higher drop-out rates (Jadidi et al. 2017) and have a harder time being successful because of "masculine" culture, discrimination, or the handicap of lower self-confidence (Bentley and Adamson 2003). Computer science is one of the fields where gender-based occupational segregation is still strong. While 57% of all employees in the US are women, only 25% of the employees in computing are women. They earn only 18% of the bachelor's degrees in computational sciences (Lehman et al. 2016). Between 1980 and 2010, 88% of all the information technology patents were introduced by male-only teams, which shows that the technology we use is invented by a strongly male-dominated community (Ashcraft et al. 2016). This may worsen the situation, as studies show that men have an advantage over women when using tools designed by other men (Beckwith and Burnett 2004; Beckwith et al. 2006).

Measuring inequalities Studies investigating gender and racial inequalities in online communities and labor markets find that the gaps are just as prevalent and relevant online as offline (Wachs et al. 2017; Hannák et al. 2017; Thebault-Spieker et al. 2015; Ge et al. 2016). Many of these studies are concerned with discrimination based on information available on user profiles (Terrell et al. 2017), social feedback by the community as a manifestation of offline discrimination in online context and algorithms reinforcing existing gender biases

(Sweeney 2013; Sandvig et al. 2014; Marom et al. 2014). Scholars are also drawing attention to the legal aspects of discrimination and labor market protections in the online world (Barzilay and Ben-David 2016).

An online platform does not need to have a financial purpose to create or reinforce offline gaps in participation and success. Several authors have studied open-source communities and their inequalities. Previous research shows that the most frequently used online knowledge sources are often created by a small minority, because cultural and algorithmic features of the platform discourage women or other underprivileged groups from contributing and editing. Studies on Wikipedia find that women are underrepresented as editors and also as subjects of the content leading to a skewed representation of knowledge (Lam et al. 2011; Reagle and Rhue 2011; Wagner et al. 2016; Menking and Erickson 2015). Similar patterns were found on OpenStreetMaps and Google MapMaker where the features identified on digital maps catered to men's tastes, as men contribute more than women (Stephens 2013). The underrepresentation of women is more pronounced in content creation than participation. A recent study of Wikipedia (Shaw and Hargittai 2018) finds evidence of a leaky pipeline: while women are not significantly less likely to have heard of Wikipedia or visited the site, they are significantly less likely to know that the site can be edited or to have made a contribution.

Stack Overflow is itself a well-studied platform. Vasilescu et al. show that women are underrepresented in this community (Vasilescu et al. 2013). Interviews with a sample of Stack Overflow users highlight the barriers women have to greater participation. Women respondents listed the lack of awareness of some site features, the intimidating community size and their fear of lacking adequate qualifications as main barriers to participation (Ford et al. 2016). Recent work by Ford, Harkins, and Parnin finds an important effect of the gender imbalance on user activity: women are more likely to engage with a post on Stack Overflow if they see other women in the conversation (Ford et al. 2017). This finding is both promising, suggesting a potential virtuous cycle of increased engagement by women, and worrisome, as higher turnover among women could compound existing disparities.

Users on the site can collect badges, tokens awarded to users for specific actions and activity, and past work has shown that this steers and influences user behavior (Anderson et al. 2012). However, research shows that gamification does not impact women's and men's behavior in the same way. The disparate influence of gamification on men and women has been observed in educational settings, in workplace and on online platforms as well (Herzig et al. 2015; Pedro et al. 2015). In an elementary school environment, a gamified educational virtual software supporting math teaching significantly improved boys' learning motivation, but it had no effect on girls' motivation or performance (Pedro et al. 2015). One potential explanation, backed by experimental evidence, is that men are socialized to have a greater preference for competition than women (Niederle and Vesterlund 2007).

Methodologically, economists have a long history of estimating gender disparities. The Oaxaca-Blinder decomposition (Oaxaca 1973; Blinder 1973) is a widely used econometric tool to disentangle the reasons behind gender differences in various outcome variables. It has also been used to study causes of obesity in different racial groups (Sen 2014), differences in career advancement prospects of men and women (Chen et al. 2010), and the difference in labor market outcomes between agency-endorsed and independent job-seekers (Stanton and Thomas 2015). Education researchers have also used Oaxaca-Blinder to explore differences in why women and men aspire to major in computer science, and how this changes over time (Sax et al. 2017).

3 Background on Stack Overflow

In this section we describe Stack Overflow as a website and a community. Stack Overflow, founded in 2008, is the largest Q&A site for computer programming. Today, the site hosts over 16 million questions and 24 million answers, and it has a global Alexa rank of 63 (Stackoverflow traffic statistics. https://www.alexa.com/siteinfo/stackoverflow.com). Previous work on Stack Overflow has highlighted its importance to the programming community as a hub of knowledge-sharing (Vasilescu et al. 2014). According to creators, their goal was to design a free access platform serving users with a high quality knowledge base (The stack overflow age. https://www.joelonsoftware.com/2018/04/06/the-stack-overflow-age). Indeed, today programmers of any level or type turn to Stack Overflow as part of their daily routine and the site is usually among the top results in Google searches for programming related queries. In this way, the knowledge shared on Stack Overflow is reused beyond the initial exchange between question-asker and answer-giver. Its user-base also significantly overlaps with that of popular code repositories such as Github (Vasilescu et al. 2013).

Stack Overflow also has influence on hiring/recruiting in the IT sector. Users can search for employment opportunities on the site's job boards. Moreover, Stack Overflow provides opportunities for them to demonstrate credible, verifiable expertise. Indeed, IT companies and recruiters often look for Stack Overflow profiles when trying to fill positions (Fawcett 2012). This is facilitated by a recently developed resume service on the site: users can turn their profile and activity into a standardized, searchable resume, ready for inclusion in the site's database of jobseekers. In this way Stack Overflow as a platform is a becoming a significant labor market matching service.

Stack Overflow's knowledge base (namely the questions and answers that have been posted) are freely available to the public - no registration is required. In order to create content, however, users have to sign up using an email address or social media account. Every account has an associated profile page which tracks a user's activity history and accomplishments on the site. Users can also enhance their profiles with biographical information, contact information, and an image. The large disparity between the number of unique monthly visitors (estimated by Quantcast to be 50 million in March 2018) and the number of active accounts made in the history of the site (less than 10 million as of March 2018, with far fewer active accounts), indicates that the vast majority of the site's users are passive: using the site's knowledge without making contributions of their own.

Registered members can post questions and answers, vote on or edit the questions and answers of other users, and interact with posts using comments. The up and down voting functionality serves as a natural content moderation, users can boost useful questions and answers and subsequent visitors have an easier time finding them. There are also elected official moderators among the community members, who can delete, modify content, and merge repeated questions into one topic.

Aside from the potential for its open and public-facing nature, Stack Overflow also has a gamification aspect (Anderson et al. 2012). Specifically, participation leads to users earning various reputation points and badges. Reputation points are primarily received for upvoted questions and answers. Users receive some moderation privileges when they accumulate enough reputation points. Bronze, silver, and gold badges can be earned through a variety of activities, for example for receiving some number of upvotes on a question or answer, editing posts for clarity, or even for visiting the site for a number of consecutive days. Gamification is a common method used both to increase user engagement and to steer users

Table 1 How users gain or lose reputation points on Stack Overflow

Outcome	Reputation
Answer Upvoted	+10
Answer Downvoted	−2, (−1 to downvoter)
Answer Accepted	+15, (+2 to acceptor)
Question Upvoted	+5
Question Downvoted	−2
Offer Bounty	−Bounty Value
Answer Wins Bounty	+Bounty Value
Answer Marked Spam	−100
Edit Accepted	+2 (max 1000/user)

towards specific behaviors deemed to benefit the community. Indeed, previous studies show that the badge system of SO has a motivating effect on the community. However, as past research indicates that men may respond more to gamification than women, this may exacerbate gender inequalities in participation (Pedro et al. 2015; Herzig et al. 2015). We note that a user's reputation and badge counts are immediately visible next to any question they ask or answer - offering a signal of the user's presence and participation on the site to others.

4 Data and Methods

In this section, we present a more detailed overview on how the Stack Overflow website works and the data we gathered about users. We also give a thorough outline of the features we extracted or created and that will serve the basis of the upcoming analyses. We now present our data collection and labeling methodology. Additionally, we introduce our dataset, focusing specifically on how the data breaks down along gender lines.

4.1 The Website

Users who create accounts on Stack Overflow can ask and answer questions as well as comment on questions or answers. For easy navigation between question and topics, users label questions with tags, indicating the topic of the question (for example if the question is about a specific programming language or algorithm). They gain reputation points, our fundamental measure of success, by receiving explicit positive feedback called "upvotes" on their questions or answers. We outline the ways users accumulate reputation in Table 1. Any user who accrues 15 reputation points gains the ability to upvote questions and answers[2] Stack Overflow also rewards specific behaviors with badges, which are tokens given for some kind of accomplishment (for example visiting the site every day for an extended period of time, receiving a set number of upvotes for a question they ask etc.).

We use the Stack Overflow API[3] to collect information on all users with at least 100 reputation points, as these users can be considered active on the website (they are granted the basic rights to comment, upvote, flag and edit on Stack Overflow). In all, we collected data on 565,171 users. To supplement the information provided by the API, we scraped data

[2] https://stackoverflow.com/help/privileges/vote-up
[3] https://api.stackexchange.com/docs

on users' activity, including the badges they collected, the tags they used, and the count of questions and answers they posted.

4.2 Feature Creation

Several of the features we use in our analysis can be extracted directly from user profiles. First, we note users' meta data, including their biography text, sign-up date, and whether they link to a personal website or social networking accounts such as Twitter, Linkedin, or Github. We operationalize these features in our models as a self-promotion index which takes a value between 0 and 1 depending on the proportion of self-promotion fields that the user has filled out. Within the biography field we check the text for the substrings "senior", "lead", "head", and "manage", assigning a dummy to each user taking the value 1 if they list any of these leadership or senior position indicators in their bio.

Second we quantify their activity on the site by how many questions they ask and answer, how often they edit posts, how many upvotes and downvotes they cast, and how often they make posts with which tags. Finally, we have information about user's outcomes and success on the site from their reputation scores and the number and types of badges they receive.

Gender inference Inferring gender of individuals from their online profiles is a complex problem. We apply a two-step approach to infer user gender, first using *genderComputer* (Vasilescu et al. 2014), a tool specifically created to infer gender of Stack Overflow users from their given usernames and location. GenderComputer considers a variety of string manipulations (for example reversing "Nohj" to get "John") to expand the scope of the inference. Location can provide additional accuracy by distinguishing, for example, between an Andrea from the UK (likely a woman) and one from Italy (likely a man). This method classified the users from our sample into 238,150 male, 24,717 female, and 302,304 unidentified users. In order to evaluate the quality of this classification, we manually examined 100 users classified as men and 100 classified as women. We found that while the method performed very well on men (97% agreement with our manual check), our manual check agreed only in 44 out of 100 cases of women. This replicates the recent finding by Ford et al. (2017) that genderComputer sacrifices precision for greater recall when inferring women users.

The second step of our inference seeks to correct this bias by applying a more conservative method based only on first names and location called *Gender Guesser*. By considering only users rated as likely male or likely female by both methods, we are left with a smaller but more accurate sample. 10,571 users are rated as highly likely women by both methods. We randomly choose 10,571 likely men (again classified as such by both methods) to obtain a balanced sample. We repeated our manual check of a random sample of accounts finding 96% agreement with our classification of men, and 84% agreement with our classification of women. This ensemble approach resembles Ford et al.'s modification of genderComputer to focus on the detection of first names within the username (Ford et al. 2017).

We acknowledge several limitations and drawbacks to our approach to inference. First, we make the simplifying assumption that gender is binary. We argue that this is a fundamental limitation of examining questions about gender differences using harvested data. Second, discarding alias usernames builds on the assumption that men and women are equally likely to adopt user names that can be mapped to their respective gender, and that this mapping does not substantially impact our hypotheses. However, previous research has shown that anonymity impacts behavior (Robertson et al. 2017), and it is possible that some users utilized an anonymous name in order to establish an independent identity. Such name selection is highlighted by the literature on gender swapping in online communities (Bruckman 1996;

Szell and Thurner 2013), where, for example, women may pose as men if they feel that they will be taken more seriously or to avoid harassment. We also note the limitations of the geographic component of our inference: a minority of users include location data, and a given location may not reflect a user's origin (for example if an Italian man named Andrea moved to the UK). Finally, name-gender databases have been shown to have significantly less accuracy when used to infer gender for non-European names (Karimi et al. 2016).

Despite these limitations, we argue that our focus on identifiable names provides the best possible data to test our hypotheses of gender behavioral and outcome differences on Stack Overflow. By limiting our dataset to users where we are highly confident about our gender inference, we gain greater confidence in the estimates of our econometric models. Moreover, our analysis includes robustness checks with 5, 10, 20, and 50 percent of our gender labels in the balanced sample randomly shuffled. These test help us better understand the effect of potential classification errors on our results. See details in Section 5.2.

Detecting user communities Given the size of the site and the diversity of topics that its users discuss, we consider that coherent communities of users may exist with significantly different patterns of behavior, norms, and outcomes for men and women. For example, users active in a more diverse community may be less likely to leave the site (Vasilescu et al. 2015), while women encountering other women are more likely to engage in a thread (Ford et al. 2017). Using a similar approach to Bosu et al. (2013), we grouped users in communities by building a network where two users are connected if their posts often share the same tags. Specifically, we created a similarity measure between users by calculating a weighted Jaccard similarity measure, defined as

$$s(u, v) = \frac{\sum_{t \in T} \min(t_u, t_v)}{\sum_{t \in T} \max(t_u, t_v)}$$

where T is the collection of all tags used at least 200 times, and t_u denotes the number of times user u made a post with tag t. We then filtered the edges using Serrano's disparity filter (Serrano et al. 2009), which, for each node, checks the weights on all its adjacent links against the null hypothesis that they are uniformly distributed. Each observed weight then has a p-value. We filter edges using this p-value ($p<.01$). The resulting network has

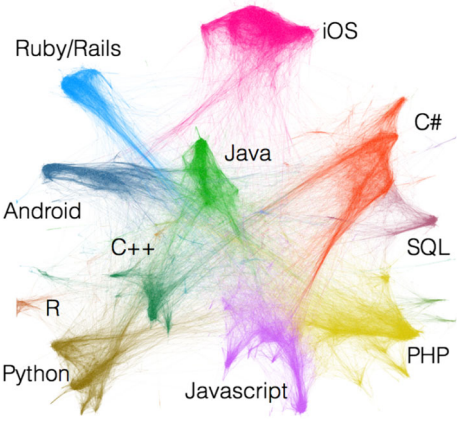

Fig. 1 User-user tag similarity network. Two users are connected if they have statistically similar tag-use patterns. Colors indicate communities detected using a community detection algorithm. Labels correspond to the most frequently used tag in each community

Table 2 Descriptive statistics of the 10 largest user communities based on a network of tag-use similarity among our balanced sample of users

Description	# of users	% male	% downvotes	% Rep. last year
C#/asp.net	2900	54%	5.5%	6%
Java	2605	49%	5.6%	7.2%
PHP	1941	51%	7.4%	6.5%
Android	1856	45%	5.5%	8.3%
Python	1665	49%	5.1%	9.3%
iOS	1548	49%	5%	6.9%
Javascript	1526	48%	7%	8.9%
C++	1390	51%	5.8%	5.9%
Angular/Node	886	51%	5.3%	14.3%
Ruby/Rails	873	55%	4.2%	6.8%

We describe each user community by interpreting the most frequently used tags in posts by its members. The last column refers to the share of the community's total reputation gained in the last year

approximately 150,000 edges connecting the roughly 21,000 users. We use the Louvain algorithm (Blondel et al. 2008) to detect communities in this network. We tune the method's resolution parameter to find larger communities to facilitate a qualitative understanding of the communities found.[4] We plot the network, visualized using a force-layout algorithm, in Fig. 1. The nodes are colored by community membership.

We manually checked the most commonly used tags in each community and found many clearly interpretable communities. The prominent programming languages and frameworks we identify in the largest communities coincide with those found in other analyses of programming language use, for instance on GitHub (Celinska and Kopczynski 2017). We describe the 10 largest communities, accounting for 80% of our users, in Table 2. Note that we sampled the males to achieve a 50-50 male-female ratio in our dataset. We see small, occasionally statistically significant gender differences. We find that the C#/asp.net, a Microsoft-developed software framework, and Ruby/Rails, a web development framework, communities have the highest representation of men, while Android, a programming language for mobile phone applications, has more women. We find that Ruby/Rails is the community with the lowest incidence of downvoting.

As past work indicates, community structure has a significant impact on user behavior and the possibilities for gaining reputation (Bosu et al. 2013). For instance, it may be easier to ask a new question or post answers in a newer community, for example on Angular/Node related questions, than in a long established community such as on C++. Therefore subsequent models explaining gender differences (see Results section) include fixed effects for user community. We also control for the size of the community, the percentage of the community that is male, and the percent of reputation generated by users in the community in the last year as a proxy for how new the community is.

[4]The modularity score, a measure of the overall quality of a partition, does not significantly change when we tune the algorithm for this purpose.

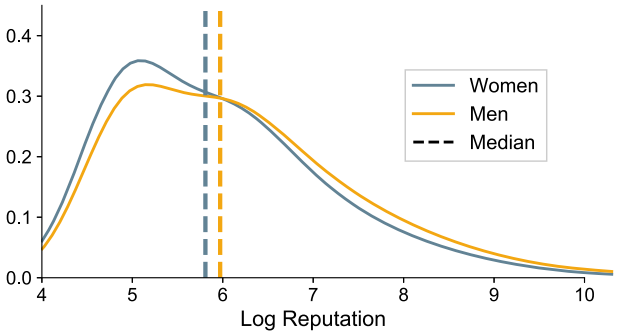

Fig. 2 Kernel density estimates of the logged reputation scores of men and women

5 Results

5.1 Descriptive Statistics: Men vs Women

The first question we ask is whether we can see a difference in the outcome measures of men and women on the site. Our key dependent variable is reputation, and we see that there are significant differences between men and women. The average reputation score is 1703 for men and 942 for women. In other words, women have on average 55% of the reputation of men. The median woman has 73% of the reputation of the median man, suggesting that many of the top reputation earners are men. The log-transformed reputation group averages are 6.1 for males and 5.8 for females, corresponding to the geometric means of 461 and 332, respectively. All differences are statistically significant (using a Mann-Whitney U test, $p < .001$). We plot the densities of the log reputation scores of men and women in Fig. 2.

We also note differences in average activity levels, outlined in Table 3. In contrast to a 2012 study of men and women on Stack Overflow (Vasilescu et al. 2012), which found that men are more active on the site across all measures, we find that women are more likely to ask questions. There are several possible explanations for this finding, for instance that the patterns of behavior on the site have changed, or because of differences in our approach to gender inference (i.e. having a lower false-positive rate among our likely women) or data selection (i.e. considering only users with at least 100 reputation). Indeed follow up work by the same authors of the 2012 study find that when controlling for overall length of engagement, women ask more questions (Vasilescu et al. 2013).

Table 3 Average activity levels across gender. Men answer 53% more questions on average than women do, while women ask 18% more

Mean activity	Women	Men
# answers	19.5	39.8
# Questions	16.4	13.5
# Edits	9.0	10.0
# Upvotes	115.0	170.0
# Downvotes	14.6	21.9
Account Age (days)	1718	1925

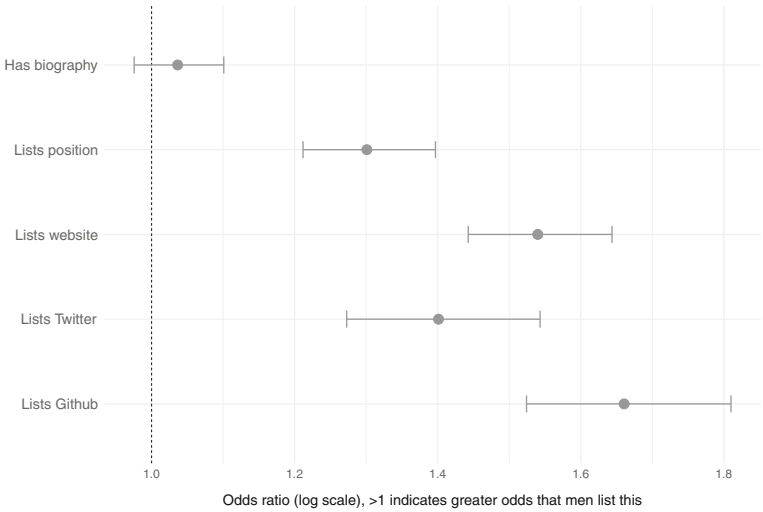

Fig. 3 Differences in self-promotion using a propensity-score matched sample of men and women. Men are significantly more likely to fill in the position, website, Twitter and Github fields. Whiskers mark 95% confidence intervals

5.2 Analysis

As outlined in the previous section we find several differences in both activity and outcome between men and women. How does the former impact the latter? We introduce a series of possible explanations for this difference, and check to see if controlling for these confounds in a regression framework can reduce or eliminate the gender gap.

First we examine the differences between how men and women share information about themselves on the site. Similar to previous findings on LinkedIn (Altenburger et al. 2017), we find that men are significantly more likely to fill out their biography, to link to their Github, LinkedIn, Twitter, or personal websites on Stack Overflow. We plot the matched log-odds ratio in Fig. 3. Second, we consider differences in tenure and find that on average men have been on the site significantly longer than women. Third, we consider that women may be overrepresented in certain communities with different norms and behaviors. We find limited evidence for gender segregation across communities, but nevertheless include community fixed-effects in later modeling efforts. Fourth, we examine differences in activity. We find that men are more likely to answer questions, while women are more likely to ask questions, with both differences significant according to Mann-Whitney U tests ($p < 0.001$).

Regressions We use a linear regression framework to explore gender differences in user reputation. We examine the relationship between activity measures such as number of questions asked, number of answers given, and number of votes cast, and reputation (log-transformed), while controlling for potential confounders such as tenure. We report our findings in Table 4. In a simple model controlling only for tenure, we find that users posting 1% more answers are expected to have 0.50% more reputation. On the contrary, the impact of asking an additional question is an order of magnitude smaller. This is an inherent feature of the current scoring system, see Table 1. The coefficient on the male term is positive and significant but close to 0.

Table 4 User reputation regressed on gender, user-level activity measures, controlling for self-promotion indicators and community-level features

	Dependent variable: ln(Reputation)		
	(1)	(2)	(3)
Male	0.04[c]	0.03[c]	−0.25[c]
	(0.01)	(0.01)	(0.02)
Answers Posted (log)	0.50[c]	0.52[c]	0.45[c]
	(0.005)	(0.005)	(0.01)
Questions Posted (log)	0.08[c]	0.09[c]	0.09[c]
	(0.004)	(0.004)	(0.01)
Votes Casted	0.09[c]	0.08[c]	0.08[c]
	(0.004)	(0.004)	(0.004)
Account Age	0.49[c]	0.50[c]	0.50[c]
	(0.01)	(0.01)	(0.01)
Male × Answers Posted			0.13[c]
			(0.01)
Male × Questions Posted			−0.01
			(0.01)
Constant	0.79[c]	1.35[c]	1.43[c]
	(0.07)	(0.38)	(0.38)
Observations	21,142	21,142	21,142
Self-promotion Controls		YES	YES
Community Controls		YES	YES
Adjusted R^2	0.61	0.62	0.63
Residual Std. Error	0.74	0.74	0.73
F Statistic	6,749.74[c]	1,734.25[c]	1,617.60[c]

The gender indicator interacted with the activity measures in the third model shows that men get more reward for posting additional answers than women do

Note: [a]p<0.1; [b]p<0.05; [c]p<0.01

Next we include controls for the users's propensity to disclose information in their profile and community measures, such as the share of men in the community and the share of total reputation earned by members in the last year. This latter measure proxies the age of the community or the newness of its topics. We find similar coefficients as in the previous model.

Finally, we include interactions of the male term and both question and answer activity. The significant positive term on the interaction of gender and the number of answers posted shows that men gain more reputation for an additional answer than women. We visualize the interaction model in Fig. 4. We use an F-test to test the null hypothesis that the coefficient on the male term and its interactions in the third model are simultaneously equal to 0. The value of the F-statistic is 128 with 1 and 21119 degrees of freedom, and the test returns p-value less than 10^{-15}. We reject the null hypothesis that the regression coefficient vector is the same for both genders. The significance of the interaction term justifies the use of

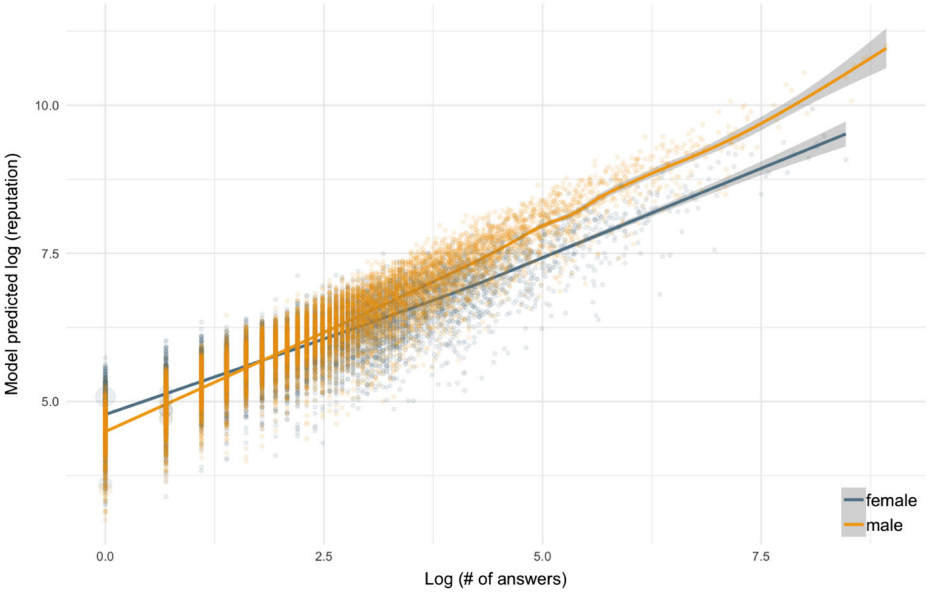

Fig. 4 We plot the marginal effect of the number of answers a user posts on his or her reputation by gender according to Model 3 in Table 4. The model controls for tenure, activity, self-promotion indicators, and community-level features. Note the difference in slopes between the genders, indicating that men get more reputation points for additional answers

a decomposition method in the following section to weigh the contributions of the various features to the overall outcome gap.

We run two robustness tests on our results. First we randomly shuffle the inferred gender on subsets of our data to test the effect of error in our classification on the observed effect. We find that the male and male $\times$ number of answers terms from Model 3 in Table 4 both remain positive and significant if we randomly shuffle 5, 10, 20, or 50 percent of the users' gender classification. For instance a 20% randomization of the gender labels shrinks the effect of the male term to -0.13 (from -0.25) and the interaction term to 0.11 (from 0.13), with both coefficients still significant at $p<.01$. The effect disappears when we completely randomize the gender labels.

Second, drawing on the observation that men are highly overrepresented among top users, we check our results dropping the top 1, 5, and 10 percent of users by reputation score. Our results are robust to this change. Finally, we combine the two robustness tests, randomly shuffling the gender label on 20 percent of our users, and dropping the top 1 percent. Both the male and interaction terms remain significant, albeit with smaller effect sizes (male coefficient: -0.07, interaction coefficient: +0.06, both significant at $p<.01$). Full model tables for the robustness tests are available on request.

Oaxaca-blinder decomposition The Oaxaca-Blinder decomposition is commonly used to measure and explain the causes of differences in averages between groups, including the wage gap between men and women (Oaxaca 1973; Blinder 1973). This decomposition for linear models allows us to observe the effect of differences in feature endowments that are used to predict the outcome between the groups (for instance that men may be on the site longer on average) and the effect of differences in how the features predict success -

and thus different coefficients - between the groups (for instance if the same increase in tenure predicts a higher reputation boost for men than woman) separately. Neumark's elaboration (1) to the method introduced a tool to examine the effect of group endowments and coefficients compared to a vector of reference coefficients (β^P) computed from the pooled OLS regression (Neumark 1988). The difference in the levels of explanatory variables weighted by the reference betas ($\triangle x\beta^P$) shows the part of the gap in the outcome variable which is explained by the differences in group averages ("explained"), while the explanatory variables weighted by the differences of the gender-specific and the pooled betas ($x^{male}(\beta^{male} - \beta^P) + x^{female}(\beta^P - \beta^{female})$) indicate the remaining, unexplained positive and negative biases. In the literature on wage differentials between groups, this second component is sometimes referred to as a measure of the discrimination present in a market. In our setting it captures the difference that would remain if women would have the same feature endowments as men.[5]

$$y^{\text{male}} - y^{\text{female}} = \triangle x\beta^P + \left[x^{\text{male}}(\beta^{\text{male}} - \beta^P) + x^{\text{female}}(\beta^P - \beta^{\text{female}})\right] \quad (1)$$

where

$$\triangle x = x^{\text{male}} - x^{\text{female}} \quad (2)$$

In our model estimating user reputation, we control for the potential explanations outlined above, including the number of answers given, questions asked, votes cast, and the age of the user account. We also include community-level features, and self-promotion dummies discussed before.

Twofold decomposition According to the twofold Oaxaca decomposition (using the pooled OLS betas as reference coefficients following Neumark (1988)), 89% of the reputation differential can be explained by the effects of differences in the explanatory variables we used (number of questions and answers posted, number of votes cast, age of the account, average reputation change in the last year within the user's tag modularity class, self-promotion and leader dummies). The difference in the effect of number of answers posted online explains 75% of the reputation differential, while the difference in the effect of account age provides an explanation for 22% of it. The remaining ("unexplained") 11% of the reputation differential might be due to gender discrimination. Given that the inclusion of more features would likely decrease the difference explained by this component, we suggest that discrimination is a limited driver of reputation inequality on the site (Fig. 5).

5.3 Discussion of the Results

We found that activity differences — mostly the difference between the amount of answers given by men and women — drive success inequality. There are a few theories in the literature than can explain the situation. Ford and her co-authors (Ford et al. 2016) found that women often hesitate to actively participate on the website because they fear they lack qualifications and because of the size and negativity of the community. While men, who are generally more competitive, thrive in this environment, women might be discouraged from answering due to these factors. We also note that the unexplained 11% of the decomposition gap may indicate that women are treated differently on the site.

[5]For a longer exposition of the equations presented here see (O'Donnell et al. 2008, p. 149-151).

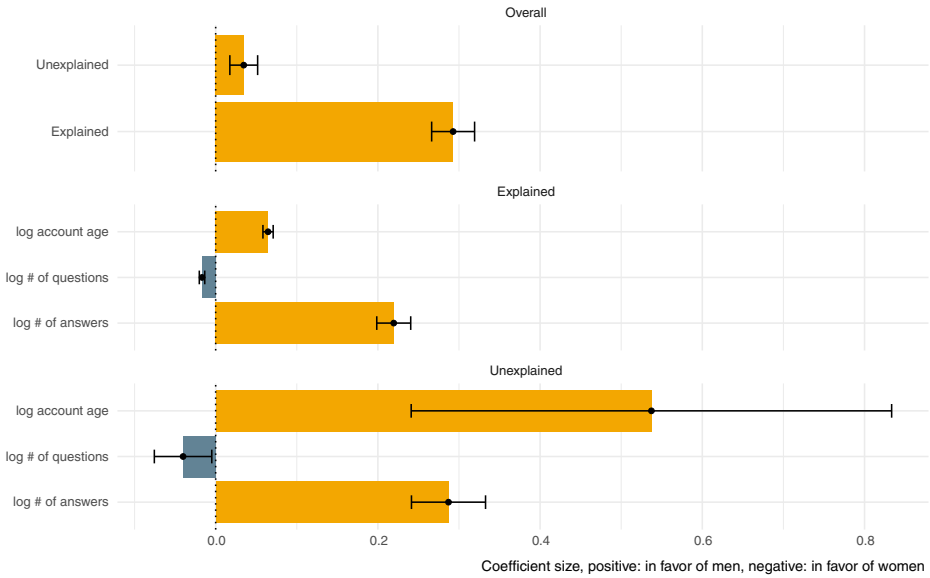

Fig. 5 The Oaxaca-Blinder decomposition of the difference in log reputation between men and women. The *Overall* subplot shows the difference decomposed into a part which is due to the known differences in endowments, accounting for 89% of the difference, and unexplained effects accounting for 11%. In the *Explained* and *Unexplained* subplots, each component is broken down into features. The difference in the amount of answers given by men and women as groups is by far the strongest explanatory factor. On all plots the whiskers indicate 95% bootstrapped confidence intervals. In the latter two plots, only selected significant terms are shown

We also see that women contribute differently to building the community's knowledge base: they are asking more questions. Stack Overflow's current system strongly incentivizes answering by rewarding upvotes on answers twice as much upvotes on questions. In the subsequent section we test how the outcome gaps would change if these rewards were equalized.

6 Proposing an Alternative Reward System

In this section we discuss one potential way to mitigate the gender differences in outcome and success on Stack Overflow: modifying the reward system. Our results suggest that differences in the rate at which men and women post answers account for the largest part of the reputation gap. On the other hand, women tend to ask more questions than men. As reputation points are almost entirely a function of the number of upvotes received on questions and answers posted, and because receiving an upvote on a question results in half of the reputation gain (+5) that receiving an upvote on an answer does (+10), we investigate what happens to the distribution of reputation scores of men and women if these rewards were equalized. In other words, we check if equalizing the rewards for good answers and good questions decreases the gender gap.

When Stack Overflow was launched in 2008, upvotes to questions and answers gave the receiver ten reputation points. In 2010 the rules were changed to their current format and reputation scores were retroactively altered for all users. In a blog post explaining the

change, one of the co-founders of the platform cited three reasons for the decision to change the system:[6]

- "We know that answers have more intrinsic value than questions, and the reputation balance should reflect that."
- "The question asker already enjoys a substantial benefit beyond reputation gain from upvotes on their question, namely, they get great answers to their question! Thus, the asker shouldn't need as much reputation gain."
- "There are a few users who ask hundreds, sometimes even thousands of questions. Over time, these users generate a fairly sizable reputation entirely through the tiny trickle of upvotes gained by these questions. In a sense, we want to discourage question asking a little bit, and make sure that people who ask questions are doing it for the right reasons and not to generate reputation."

Independent of the issue of gender disparities, we argue that the proposed change has merit when considered against these points, especially when considering the site's increased importance as a knowledge resource since 2010. We do not agree, for example, with the value judgement that answers have more intrinsic value than questions: without the question there would be no answer. Stack Overflow distinguishes itself from Wikipedia or textbooks as a knowledge resource by providing applicable answers to real-world user-generated questions. The service of asking a genuine question, the answer to which may seem simple to an expert, is part of what gives Stack Overflow its appeal above and beyond the example cookbook solutions available in many programming references or textbooks.We also note that a single question can generate multiple useful answers - suggesting that any one person answering a question can still learn from other answers. Finally, improvements in site moderation and semi-automated detection of duplicate questions (Zhang et al. 2015) likely reduces the prevalence of reputation mining by asking repetitive questions.

Increasing rewards to good questions may help to make the site more inclusive by offering a less competitive and speed-oriented way to build one's reputation. For example, recent work on user strategies for gaining reputation on the platform focuses entirely on answers, finding that answering questions quickly (ideally first) is one of the best ways to quickly collect reputation (Bosu et al. 2013). The authors also find that focusing on areas where there are fewer experts, or answering questions on off-peak hours are productive strategies. Such time pressures do not play the same role when one is asking a question.

To calculate the revised reputation score, we collect additional data on each user's activity. We calculate the revised reputation scores of all users by counting question upvotes as being worth 10 points. Recall that women have on average 55% of the reputation of men, and the median woman has 73% of the reputation of the median man. Using the revised reputation, women have 71% of the reputation of men on average, and the median female has **16% more** reputation than the median male. We see this shift in Fig. 6, where we plot the distribution of log reputation and log revised reputation by gender. Indeed women have much lower variance: the low and high end of the distribution are proportionally much more male.

If additional reputation for asking questions even slightly increases the engagement of some women on Stack Overflow, peer effects might encourage other women to post (Ford et al. 2017). Yet we emphasize that this potential change is only one strategy to address gender disparities on Stack Overflow. Indeed, one should also be interested in why men are so much more likely to give answers, and how one might encourage women to give

[6] https://stackoverflow.blog/2010/03/19/important-reputation-rule-changes/

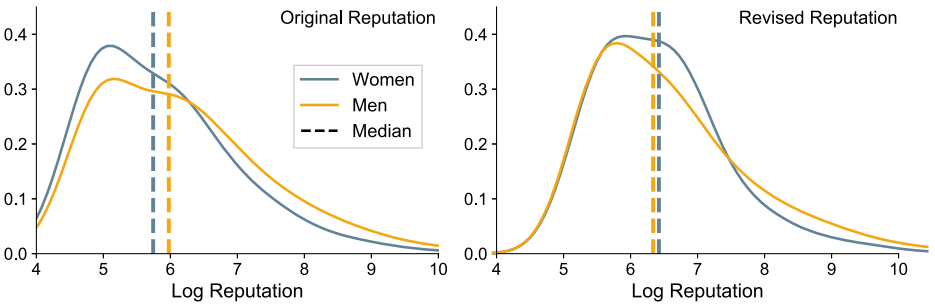

Fig. 6 Distributions of (log) reputation and revised reputation for men and women. Dotted vertical lines indicate group medians. Note that although women have higher median in the revised reputation plot, men still have higher average reputation as evidenced by their overrepresentation in the long-tail of success

more answers. Furthermore, one must consider how all users would change their behavior if the scoring system stopped favoring answers. As men might respond more readily to gamification, it is possible that they would begin to ask more questions.

In order to understand why the revised reputation scores do not change trends among the most successful users, we investigated how users of different reputation levels contribute to the site. Figure 7 shows the contribution rates broken down by activity type for each reputation decile. While questions are relatively evenly distributed across the groups, most answers are given by a small number of "experts". More precisely, 68% of answers are given by the top decile of users, while only 28% of questions are. While this result is somewhat intuitive given that the motivation for asking a question is a lack of knowledge, we think it brings attention to an important issue. Increasing rewards for questions encourages users from a broader spectrum of the population, most critically the learners who form an essential part of the site.

7 Discussion

In this paper we investigated gender differences in the activity and success of users on Stack Overflow. Women are highly underrepresented in the data we collected, not only compared

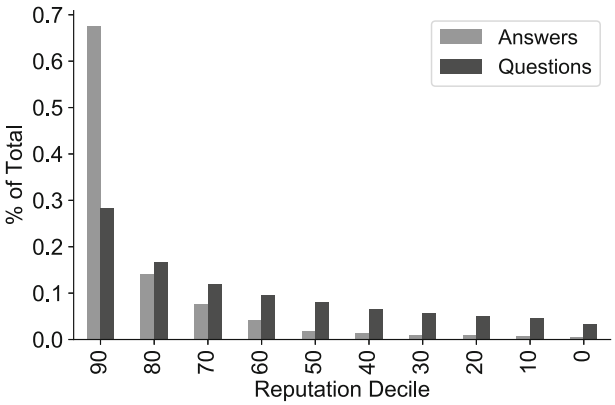

Fig. 7 Share of total answers and questions posted by users group by reputation decile

to male users of the website but also compared to their presence in the IT labor market more broadly. Moreover, according to our success measures they are also less successful on the site than men. These differences can partly be explained by different participation or activity behavior. While women ask more questions than men, men are more active in all other forms of activity. Given that Stack Overflow's current reward system favors popular answers over popular questions, these differences in activity leads to large differences in outcome. Excluding this effect and controlling for important user and community level explanatory features, only 11% of the reputation gap, which may be due to "perceptional" discrimination (among other confounders), remains unexplained. Given our finding that the current rewards system favors "male behavior", we propose a new reputation system which equalizes the reward for upvotes on questions and answers. We find that the new scoring system reduces differences in the group means, and in fact leaves the median woman with slightly higher reputation than the median male.

Limitations Our analysis has several limitations. First, we emphasize that our data was selected from individuals who registered and managed to accumulate at least 100 reputation points. As research on gender inequalities has shown again and again, survival bias likely influences the features of women in our data. Hence we must acknowledge that the men and women in our data form a biased sample of all users of the site. The individuals who are active on the site are likely to be among those who are more able to take risks, more likely to prefer competition, and less vulnerable to harassment. This would lead to an underestimation of gender differences in behavior and outcome on the site. Moreover, a large part of Stack Overflow's audience is silent. Statistics from the site suggest that there are tens of millions of monthly visitors but only half a million accounts with 100 reputation points. It is likely that these roughly 500,000 users are less than 10% of people who have used the site for learning. The true motivations for joining or staying silent is a very interesting question and could be addressed by gathering qualitative data (Ford et al. 2016).

We also reiterate the issues we raised about inferring gender from online data. We may underestimate the participation of women by missing those who pose anonymously or as men. We also acknowledge a western bias in the method we used to detect gender: studies show that such methods suffer from significantly higher error rates on non-western names (Karimi et al. 2016).

Another drawback of our analysis is that we have no information on the quality of the questions and answers posted outside the user-level upvote and downvote scores. Since quality likely interacts with the audience's feedback and the existing reputation of users these are potentially important missing controls.

Lastly, we do not know the effect of current success on future success; in other words, whether the rich get richer. Research on other platforms shows evidence for such a reinforcement effect (Muchnik et al. 2013). Our snapshot only allows us to hypothesize that such an effect may be present. More importantly, the limitations of our work highlight the need for a greater understanding of how success on Stack Overflow impacts success in the job market. We believe that these limitations can be addressed in future work by integrating experimental methods, user surveys, and the collection of longitudinal data about users' career paths on the site.

Impact We believe that we are at a crucial moment in the process of inclusion of women into the STEM and IT labor markets. One positive indicator of progress on the site is that women are more active question posters than men, while in a 2012 study men were significantly more active than women in all forms of activity (Vasilescu et al. 2012). This is

perhaps a result of an increased focus of the Stack Overflow developers on improving the diversity and inclusiveness of their platforms (Ford et al. 2017), and is in line with changes observed in the annual Stack Overflow survey.[7] However it is not enough to encourage women to start learning, and it is important to continue efforts to support and include a diverse user population reflective of the broader population.

Though our findings and subsequent recommendation address a specific gender gap found on the site we analyze, we argue that there is more work to be done on why male and female behavior differs so significantly in the first place. The more the difference is due to the diversity of experience or perspective between the groups, the more our recommendation is a useful solution to gender gap in participation. Better recognition of the validity of alternative behaviors would likely encourage participation. On the other hand, given the literature on gendered barriers to participation in IT, we suspect that a large part of the behavioral difference is a legacy of these barriers and our recommendation is only a partial solution. In other words, if women are answering fewer questions because they have been discouraged from speaking up, fear harassment, or lack self-confidence, one cannot solve the overarching problem by increasing the rewards to questions.

One promising recent development on Stack Overflow highlighting the need for more targeted intervention is the launch of a mentoring program for new users (Ford et al. 2018). New users asking questions entered "Help Rooms" to obtain constructive criticism on their questions before posting on the main site. Mentored users asked higher quality questions and were more likely to feel part of the community on Stack Overflow than control group users. In parallel with considering revisions to site design, more work is needed on how to scale this kind of intervention.

More broadly, we highlight the importance of auditing systems and evaluating the algorithms of open-source communities. Even though these systems have platforms that are more transparent, owners that are more benign, and financial impacts that are lower than large corporations like Google or Facebook, their social and labor market impacts can still be very large. As such, their owners share responsibility for the outcomes of women in the IT community, and the culture of open-learning more generally.

Acknowledgements Open access funding provided by Vienna University of Economics and Business (WU). We would like thank Agnes Horvát, Daniel Larremore, Piotr Sapiezynski, Kenny Joseph, and participants of Central European University's Gendered Creative Teams Workshop for helpful comments and insights. We thank Kristen Altenburger and Dorota Celinska for advice regarding matching and the Oaxaca-Blinder decomposition, respectively. We also acknowledge the comments of anonymous referees.

[7] https://insights.stackoverflow.com/survey/2017

References

Ahuja MK (2002) Women in the information technology profession: a literature review, synthesis and research agenda. Eur J Inf Syst 11(1):20–34

Altenburger K, De R, Frazier K, Avteniev N, Hamilton J (2017) Are there gender differences in professional self-promotionan empirical case study of linkedin profiles among recent mba graduates. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15615

Anderson A, Huttenlocher D, Kleinberg J, Leskovec J (2012) Discovering value from community activity on focused question answering sites: a case study of stack overflow. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 850–858. ACM

Ashcraft C, McLain B, Eger E (2016) Women in tech: The facts

Barzilay AR, Ben-David A (2016) Platform inequality: gender in the gig-economy. Seton Hall L Rev 47:393

Beckwith L, Burnett M (2004) Gender: an important factor in end-user programming environments? In: 2004 IEEE symposium on visual languages and human centric computing, pp 107–114. IEEE

Beckwith L, Kissinger C, Burnett M, Wiedenbeck S, Lawrance J, Blackwell A, Cook C (2006) Tinkering and gender in end-user programmers' debugging. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 231–240. ACM

Bentley JT, Adamson R (2003) Gender differences in the careers of academic scientists and engineers: a literature review. Special report

Blau FD, Kahn LM (2016) The gender wage gap: extent, trends, and explanations. Tech. rep., National bureau of economic research

Blinder AS (1973) Wage discrimination: reduced form and structural estimates. J Hum Resour 8:436–455

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 2008(10):P10008

Bosu A, Corley CS, Heaton D, Chatterji D, Carver JC, Kraft NA (2013) Building reputation in stackoverflow: an empirical investigation. In: 2013 10th IEEE working conference on mining software repositories (MSR), pp 89–92. IEEE

Bruckman A (1996) Gender swapping on the internet. High noon on the electronic frontier: conceptual issues in cyberspace pp 317–326

Celinska D, Kopczynski E (2017) Programming languages in github: a visualization in hyperbolic plane. In: ICWSM, pp 727–728

Chen Z, Roy K, Gotway Crawford CA (2010) Examining the role of gender in career advancement at the centers for disease control and prevention. Am J Public Health 100(3):426–434

Clayton D, Lynch T (2002) Ten years of strategies to increase participation of women in computing programs: the central queensland university experience: 1999–2001. ACM SIGCSE Bulletin 34(2):89–93

Cohoon JM, Aspray W (eds) (2006) Women and information technology: Research on underrepresentation, 1st edn., vol 1. The MIT Press, Cambridge. https://EconPapers.repec.org/RePEc:mtp:titles:0262033453

Fawcett H (2012) 3 unusual q&a sites to source it talent from - quora, github, stackoverflow. https://www.socialtalent.com/blog/technology-2/3-unusual-qa-sites-to-source-it-talent-from-quora-github-stackoverflow

Ford D, Harkins A, Parnin C (2017) Someone like me: how does peer parity influence participation of women on stack overflow? In: 2017 IEEE symposium on visual languages and human-centric computing (VL/HCC), pp 239–243. IEEE

Ford D, Lustig K, Banks J, Parnin C (2018) We don't do that here: how collaborative editing with mentors improves engagement in social q&a communities. In: Proceedings of the 2018 CHI conference on human factors in computing systems, p. 608. ACM

Ford D, Smith J, Guo PJ, Parnin C (2016) Proceedings of the 2016 24th ACM SIGSOFT international symposium on foundations of software engineering, pp 846–857. ACM

Ford H (2016) How wikipedia's silent coup ousted our traditional sources of knowledge. https://makebuildplay.log/2016/11/02/how-wikipedias-silent-coup-ousted-our-traditional-sources-of-knowledge/

Ge Y, Knittel CR, MacKenzie D, Zoepf S (2016) Racial and gender discrimination in transportation network companies. Working Paper 22776, National Bureau of Economic Research. http://www.nber.org/papers/w22776

Glass B (2016) Beyond point and click. http://burning-glass.com/research/coding-skills/

Hannák A, Wagner C, Garcia D, Mislove A, Strohmaier M, Wilson C (2017) Bias in Online Freelance Marketplaces: evidence from TaskRabbit and Fiverr. In: 20th ACM conference on computer-supported cooperative work and social computing (CSCW 2017). Portland, OR

Herzig P, Ameling M, Schill A (2015) Workplace psychology and gamification: theory and application. In: Gamification in education and business, pp 451–471. Springer

Horvath JA (2014) Inside the github scandal: is sexism part of the valley's dna? https://www.theverge.com/2014/3/19/5526574/github-sexism-scandal-julie-ann-horvath

Jadidi M, Karimi F, Wagner C (2017) Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. arXiv:1704.05801

Karimi F, Wagner C, Lemmerich F, Jadidi M, Strohmaier M (2016) Inferring gender from names on the web: a comparative evaluation of gender detection methods. WWW '16 Companion

Lam STK, Uduwage A, Dong Z, Sen S, Musicant DR, Terveen L, Riedl J (2011) WP:Clubhouse? An exploration of wikipedia's gender imbalance. In: Proc. of WikiSym

Lehman KJ, Sax LJ, Zimmerman HB (2016) Women planning to major in computer science: who are they and what makes them unique. Comput Sci Educ 26(4):277–298

Lerner J, Tirole J (2002) Some simple economics of open-source. The Simple Economics of Open Source 50:197–234

Leslie SJ, Cimpian A, Meyer M, Freeland E (2015) Expectations of brilliance underlie gender distributions across academic disciplines. Science 347(6219):262–265

Marom D, Robb A, Sade O (2014) Gender dynamics in crowdfunding (kickstarter). SSRN Working Paper 2442954(430):1–75

Menking A, Erickson I (2015) The heart work of wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp 207–210. ACM

Muchnik L, Aral S, Taylor SJ (2013) Social influence bias: a randomized experiment. Science 341(6146):647–651

Neumark D (1988) Employers' discriminatory behavior and the estimation of wage discrimination. J Hum Resour 23:279–295

Niederle M, Vesterlund L (2007) Do women shy away from competition? do men compete too much Q J Econ 122(3):1067–1101

Oaxaca R (1973) Male-female wage differentials in urban labor markets. Int Econ Rev 14:693–709

O'Donnell O, Van Doorslaer E, Wagstaff A, Lindelow M (2008) Analyzing health equity using household survey data: a guide to techniques and their implementation. World Bank, Washington

Pedro LZ, Lopes AM, Prates BG, Vassileva J, Isotani S (2015) Does gamification work for boys and girls?: An exploratory study with a virtual learning environment. In: Proceedings of the 30th annual ACM symposium on applied computing, pp 214–219. ACM

Reagle J, Rhue L (2011) Gender bias in wikipedia and britannica. Int J Commun 5:1138–1158

Republic T (2014) The state of women in technology. http://www.techrepublic.com/article/the-state-of-women-in-technology-15-data-points-you-should-know/

Robertson RE, Tran FW, Lewark LN, Epstein R (2017) Estimates of non-heterosexual prevalence: the roles of anonymity and privacy in survey methodology. Arch Sex Behav 47:1–16. https://doi.org/10.1007/s10508-017-1044-z

Sandvig C, Hamilton K, Karahalios K, Langbort C (2014) Auditing algorithms: research methods for detecting discrimination on internet platforms. In: Proceedings of "Data and Discrimination: Converting Critical Concerns into Productive Inquiry", a preconference at the 64th Annual Meeting of the International Communication Association

Sax LJ, Lehman KJ, Jacobs JA, Kanny MA, Lim G, Monje-Paulson L, Zimmerman HB (2017) Anatomy of an enduring gender gap: The evolution of women?s participation in computer science. J High Educ 88(2):258–293

Sen B (2014) Using the oaxaca–blinder decomposition as an empirical tool to analyze racial disparities in obesity. Obesity 22(7):1750–1755

Serrano MÁ, Boguná M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci 106(16):6483–6488

Shaw A, Hargittai E (2018) The pipeline of online participation inequalities: the case of wikipedia editing. J Commun 68(1):143–168

Stanton CT, Thomas C (2015) Landing the first job: the value of intermediaries in online hiring. Rev Econ Stud 83(2):810–854

Stephens M (2013) Gender and the geoweb: divisions in the production of user-generated cartographic information. GeoJournal 78(6):981–996

Sweeney L (2013) Discrimination in online ad delivery. http://ssrn.com/abstract=2208240

Szell M, Thurner S (2013) How women organize social networks different from men, vol 3, Sci Report

Terrell J, Kofink A, Middleton J, Rainear C, Murphy-Hill E, Parnin C, Stallings J (2017) Gender differences and bias in open source: Pull request acceptance of women versus men. PeerJ Comput Sci 3: e111

Thebault-Spieker J, Terveen LG, Hecht B (2015) Avoiding the south side and the suburbs: the geography of mobile crowdsourcing markets. In: Proceedings of the 18th ACM conference on computer supported cooperative work and social computing

Vasilescu B, Capiluppi A, Serebrenik A (2012) Gender, representation and online participation: a quantitative study of stackoverflow. In: 2012 international conference on social informatics (SocialInformatics), pp 332–338. IEEE

Vasilescu B, Capiluppi A, Serebrenik A (2013) Gender, representation and online participation: a quantitative study. Interacting with Comput 26:iwt047

Vasilescu B, Capiluppi A, Serebrenik A (2014) Gender, representation and online participation: a quantitative study. Interacting with Computers 26(5):488–511. https://doi.org/10.1093/iwc/iwt047

Vasilescu B, Filkov V, Serebrenik A (2013) Stackoverflow and github: associations between software development and crowdsourced knowledge. In: 2013 international conference on social computing (SocialCom), pp 188–195. IEEE

Vasilescu B, Posnett D, Ray B, van den Brand MG, Serebrenik A, Devanbu P, Filkov V (2015) Gender and tenure diversity in github teams. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp 3789–3798. ACM

Vasilescu B, Serebrenik A, Devanbu P, Filkov V (2014) How social q&a sites are changing knowledge sharing in open source software communities. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, pp 342–354. ACM

Wachs J, Hannák A, Vörös A, Daróczy B (2017) Why do men get more attention? Exploring factors behind success in an online design community. In: 11th international AAAI conference on web and social media, ICWSM '17

Wagner C, Graells-Garrido E, Garcia D, Menczer F (2016) Women through the glass ceiling: gender asymmetries in wikipedia. EPJ Data Sci 5(1):5

Wired (2014) Why free online classes are still the future of education. https://www.wired.com/2014/09/free-online-classes-still-future-education/

Zhang Y, Lo D, Xia X, Sun JL (2015) Multi-factor duplicate question detection in stack overflow. J Comput Sci Technol 30(5):981–997

Anna May holds a master's degree in Economic Policy from Central European University, Budapest. Her research interest lies in gender inequalities on online platforms, her thesis is about gender differences in entrepreneurial ambitions and success. Currently, she is working as a data scientist.

Johannes Wachs is a PhD Candidate in the Department of Network and Data Science at Central European University. He is a computational social scientist specializing in network methods. His dissertation studies corruption and collusion in public procurement markets. He is also interested in quantifying bias and inequalities on the web. In 2019 he will be a postdoc at the Chair for Computational Social Sciences and Humanities at RWTH Aachen.

Anikó Hannák is an assistant professor at the Vienna University of Economics and Business, and faculty member of the Complexity Science Hub. Her main interest lies in computational social sciences. She is focusing on the co-evolution of online systems and their users with a focus on algorithmically aided online platforms. Since big data algorithms learn on human data, they are likely to pick up on social biases and unintentionally reinforce them. In her PhD work, Aniko created a methodology called "algorithmic auditing", which tries to uncover the potential negative impacts of large online systems. Examples of such audits include examining the filter bubble effect on Google Search, online price discrimination or detecting inequalities in online labor markets.