Gender, Representation and Online Participation: A Quantitative Study of StackOverflow

Bogdan Vasilescu*
Eindhoven University of Technology,
Eindhoven, The Netherlands
b.n.vasilescu@tue.nl

Andrea Capiluppi
Brunel University
London, United Kingdom
andrea.capiluppi@brunel.ac.uk

Alexander Serebrenik†
Eindhoven University of Technology,
Eindhoven, The Netherlands
a.serebrenik@tue.nl

***Abstract*—Online communities are flourishing as social meeting web-spaces for users and peer community members. Different online communities require different levels of competence for participants to join, and scattered evidence suggests that women can be overly under-represented. Moreover, anecdotal evidence of the Q&A website StackOverflow suggests that women withdraw from unfriendly online communities.**

Due to the lack of empirical evidence on the matter, this paper provides a quantitative study of the phenomenon, in order to assess the representation and social impact of gender in StackOverflow. This study positions itself within recent and focused international initiatives, launched by the European Commission in order to encourage women in the field of sciences and technology. Our findings confirm that men represent the vast majority of contributors to StackOverflow. Moreover, men participate more, earn more reputation, and engage in the "game" more than women do.

I. Introduction

Online communities and social sites represent an extension of the very well known "open source" phenomenon: individuals use their own free time to gather around a common web space, to discuss, socialize or request support from other contributors. Current online communities target a wide spectrum of diverse users, ranging from the general audience (e.g., Facebook), professionals (e.g., LinkedIn), or IT experts specifically (e.g., StackOverflow). The purposes of these sites can also be very diverse: some provide the ability to share content such as source code fragments (e.g., snipplr), entire projects (e.g., Github, bitbucket) or images (e.g., Flickr). Others support knowledge sharing by means of questions and answers (e.g., StackOverflow) or news postings (e.g., reddit).

The gender representation in Science, Technology, Engineering, and Mathematics (STEM) related subjects raises significant attention of researchers and academics [7], [11], [17], [24], [29], as well as of policy-makers [1], all noting a significant under-representation of women. The main reasons behind such under-representation have been studied mostly qualitatively, to delineate the issue, formulate the main reasons of the imbalance between the number of female and male participants, and propose initiatives to attract more women to sciences, in terms of both majoring in STEM-related studies as well as choosing STEM-related career paths. In particular, encouraging more women to participate in Information Technology, Computer Science, and Computer Engineering is seen as having potential benefits not only to women, but also to society [41, p.235].

In addition to gender and STEM-related studies and careers in general, there is also the issue of representation of women in the use of technology and online communities. The use of Internet technologies is not as unbalanced as the access to careers and vocational studies [5]. Still, software development remains a predominantly male activity, especially for Open Source: all surveys reviewed in [10] agree that only 1-5% of the open source developers are women. This is in sharp contrast with the 28% female employees with computer and mathematical occupations reported in [26].

The focus on gender under-representation in online communities is further motivated by anecdotal observations: it has been suggested that the Q&A website StackOverflow (SO) strongly promotes oneupmanship; fosters flame-wars and the down-voting of individuals; and it is based on earning prizes, reputation and badges, that allow participants to access new features and gain more control on others' postings [36], [37]. Experience suggests that this results in a lesser participation by female users, who do not engage with the community or use gender-neutral names to be accepted by the mostly male audiences, while male users sometimes masquerade as females believing other (male) users would be less aggressive towards them and their questions. Similar "gender swapping" has been observed in an online poker community [42].

This paper is an attempt to quantitatively evaluate the presence of women in StackOverflow, and to compare their levels and duration of engagement (as compared to the male counterparts). The main rationale of this study is based on the fact that no empirical studies have been performed yet on how gender plays a role in highly-skilled software-development-related online communities, while most of the evidence remains at the anecdotal level.

This paper is organised as follows: Section II deals with the research design, questions and metrics used in the study; Section III summarizes the issues in collecting and aggregating the necessary data; Section IV presents the results of a pilot survey; Section V presents the results; Section VI reports on

*Supported by the Dutch Science Foundation project "Multi-Language Systems: Analysis and Visualization of Evolution—Analysis" (612.001.020).

†Part of this research has been carried out during the author's stay at the Université de Mons, Belgium, supported by the F.R.S-F.N.R.S. under the grant BSS-2012/V 6/5/015.

related work, Section VII identifies the threats to validity, while Section VIII concludes.

II. Research Design

This section presents the research design of this study following the *Goal-Question-Metric* (GQM) approach [4].

A. Goal

The aim of this work is to produce a comprehensive study on how and when women engage in StackOverflow.

Rationale: what is currently known about this "disengagement" phenomenon is still at the anecdotal level, and qualitative and quantitative studies are needed for two reasons: to produce a solid and reproducible understanding of the reasons; and to evaluate the possible long-term implications of such online behaviour.

B. Questions

This paper addresses the following research questions:

RQ$_1$ **What are the issues of identifying gender in online communities?**

Rationale: the identification of gender in online activities is complicated by several factors: some communities do not record the gender of their participants; users often choose gender-neutral names, or opposite-sex names to cope with a male-dominated environment; in specific countries, certain names are "unisex", therefore the resolution of names to gender has to be country-specific.

RQ$_2$ **What is the participation rate of women in StackOverflow?**

Rationale: while the sharp decline of women in STEM-related subjects is well known, a quantitative study of how many women participate in StackOverflow (or any other software-development-related online communities) has not been achieved yet. Before trying to understand the reasons behind a possible under-representation of women, it is necessary to first delineate the issue.

RQ$_3$ **What are the types of participation of women in StackOverflow?**

Rationale: even in case of extreme skewness in the representation of women in StackOverflow, it is important to define whether women and men follow similar patterns of contribution and engagement. Showing that women engage less than men in communities, but achieve similar levels of contribution would produce a picture of a (relatively) "healthy" community. On the other hand, a community with a skewed representation of gender, and where the levels of participation varies substantially with gender would suggest a gender-specific community.

C. Metrics

The representation of gender and their levels of engagement are measured using various attributes:

- The *number of women and men* participating in online activities. Participation occurs when a user proposes a new question, or attempts to answer an existing one[1];
- The *number of questions* posted by an individual to the community;
- The *number of answers* given by an individual to pending questions;
- The *length of engagement* in the community, i.e., the number of days between the first question or answer, and the latest question or answer given by a user.

Based on the questions and metrics formulated above, we posit a number of null hypotheses (reported in Table I), to be tested via statistical testing. The alternative hypotheses test whether StackOverflow is gender-specific and biased towards men, and they are drawn from the collected anecdotal evidence. The most important statistical test we apply is the Mann-Whitney test, a non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other [22]. The test consists of calculating a test value U and comparing the calculated with the distribution which is known under the null hypothesis. The result of this comparison is a p-value. If the p-value is lower than the predefined threshold (we use the traditional threshold of 0.05) than we can reject the null hypothesis and accept the alternative hypothesis.

Null (H_0)	H_1	RQ	Test
$H_{1,0}$: women and men are similarly represented	$H_{1,1}$: women are under-represented, which reflects the current trend of women enrolling in STEM-related subjects	RQ$_1$, RQ$_2$	# of women and men
$H_{2,0}$: women formulate a number of questions statistically similar to men's	$H_{2,1}$: men formulate more questions	RQ$_2$, RQ$_3$	Mann-Whitney
$H_{3,0}$: women provide a number of answers statistically similar to men's	$H_{3,1}$: men provide more answers	RQ$_2$, RQ$_3$	Mann-Whitney
$H_{4,0}$: women engage for a length of time statistically similar to men's	$H_{4,1}$: men engage for longer	RQ$_2$, RQ$_3$	Mann-Whitney
$H_{5,0}$: women and men achieve similar levels of reputation	$H_{5,1}$: men achieve larger reputation levels	RQ$_3$	Mann-Whitney

TABLE I
Null hypotheses to be tested, and their relation to the research questions

[1] Other events are possible in SO, such as commenting/editing posts. We only analyse participation related to posing/answering questions, considered the core activities for a Q&A website.

III. Empirical Approach

StackOverflow does not record gender of participants. The empirical approach followed to infer gender involves *automatic* and *manual* steps. The automatic process comprises inferring gender *based on a person's name* and, if available, their location. The manual process comprises inferring gender *based on a person's avatar picture*, or *based on additional data sources*, such as Github, Twitter, Flickr, or LinkedIn. The manual process was performed only for those participants for which the automatic process did not infer a gender. All results were manually reviewed. Details about data extraction, as well as specific challenges pertaining to the datasets are described below. The accuracy of the gender resolution process is discussed in Section IV.

A. Obtaining the data

StackOverflow[2] is a programming questions and answers (Q&A) website collaboratively built and maintained by programmers, and owned by Stack Exchange, Inc. SO uses gamification and an activity-based reputation system: users receive badges for different actions performed on SO (e.g., resurrecting and editing posts that were inactive for long periods, up voting competing answers, or sharing links to questions in order to attract more viewers); similarly, users earn reputation points by posting interesting questions and answers (as reflected by the up votes received from the SO community). The higher the reputation and the more badges one has, the more control she has over SO and other members' postings (e.g., users having earned certain badges can be elected to help moderate the site).

All public data in SO (including the list of members and data about their activity) can be downloaded as part of the Stack Exchange data dump[3]. In this paper we explore the data dump dated April 2012. This data set contains information about 1078708 registered users. Since our gender-resolution process is only partly automatic, we decided to sample a smaller set: to obtain a 2% margin error and 99% confidence, a random sample of 4,144 SO users was extracted.

B. Automatic gender resolution

A person's name is often indicative of their gender. For example, John is a common male English first name, while Claire is a common female English first name. Corroborated with location information, even more accurate inferences can be made about one's gender based on their name. For example, Andrea is a common male first name in Italy, but a common female one in Germany. To support this approach, we iteratively built lookup tables with first names for countries where this information was accessible online. Whenever available (e.g., when the data came from national statistics institutes), we also record the name usage frequency[4].

1) Preprocessing: The goal of the preprocessing step is obtaining the *(name, country)* tuples whenever possible.

We start by preprocessing the names. To aid the name-based gender resolution process, we first convert the names in Leet to Latin. For example, w35l3y is converted to Wesley.

Next, we tried to identify real names of the participants that choose not to disclose them, using *nicknames* (e.g., Carrotman, CoffeeCode) or standard SO-assigned usernames (e.g., user4106) instead. To determine the real names of such SO users we crawl and parse personal webpages, linked from the SO profile pages. Moreover, if one person has multiple SO accounts, information obtained from one of the accounts can be used to infer gender for another one. To identify accounts belonging to the same person, we make use of email hashes[5]: accounts associated with the same hash have the same email address, and, hence, belong to the same person. For example, if a user with a standard SO-assigned username shares the email hash with *George Washington*[6], so gender can be inferred using the contributor's name *George*.

Location information is available only for the SO users that choose to describe it in their profiles. However, only a fraction of the users in our sample specify location (821 out of 4,144, or 19.8%), and not all user-specified addresses refer to geographic locations (e.g., The Matrix). Therefore, the locations were parsed via the Google Maps geocoding service[7], and the relative country (if available) was recorded.

2) Gender resolution process: We developed a Python tool that resolves a name using the lookup tables discussed above, and a number of heuristics. The tool takes a *(name, country)* tuple as input, and returns one of "female", "male", or "x" (i.e., no gender can be inferred). The resolution algorithm starts with the identification of the first and the last name, and continues with gender detection based on gender-specific last name forms (e.g., -ova in Russian), country-specific lookup tables, cross-country lookup and diminutive resolution.

For example, given *(Anna Akhmatova, Russia)* the tool infers "female" due to a gender-specific last name form; for *(Andrea Mantegna, Italy)* the tool chooses "male" since Andrea is much more frequent as a male name in Italy; for *(Bogdan Lalić, Croatia)* also "male" despite the fact that we do not have data for Croatia: Bogdan is recorded only as male in all lookup tables that include this name. Observe, however, that we cannot infer gender for *(Andrea Demirović, Montenegro)* since we do not have data for Montenegro and different countries list Andrea as male or as female.

[2] http://stackoverflow.com

[3] http://www.clearbits.net/torrents/2017-apr-2012

[4] We have compiled lists for Albania, Australia*, Belgium*, Brazil, Canada*, Czech Republic, Finland, France, Greece, Hungary, India, Iran, Ireland*, Israel, Italy*, Latvia, Norway*, Poland, Romania, Russia, Slovenia*, Somalia, Spain*, Sweden*, The Netherlands, Turkey, UK*, Ukraine, USA*, and Vietnam. The asterisk denotes countries with frequency information.

[5] The actual email addresses are not publicly available, for privacy reasons; the MD5 hashes, however, are.

[6] Here and elsewhere due to privacy reasons we do not disclose usernames of the actual SO users but replace them by with names exhibiting similar patterns.

[7] https://developers.google.com/maps/documentation/geocoding/

If none of the above results in a resolved gender, and the name contains a single name part (i.e., it resembles a username), we assume it is formatted according to common naming conventions for usernames [6] (e.g., johns for John Smith), and restart the process (e.g., with john derived from johns).

C. Manual gender resolution

Typically not all participants choose names amenable for the name-based gender resolution process discussed above (e.g., because they prefer nicknames such as CoffeeCode). To improve the accuracy of the automatic gender inference process we manually inspect additional sources of information: avatar pictures (available for some of the SO users), and websites such as Github, Twitter, Flickr, or LinkedIn, which may help reveal a person's real name.

SO users have the option of displaying an avatar picture on their profile pages. We have manually inspected the avatar pictures of the SO users, and tried to infer gender. However, not all users upload pictures of themselves (e.g., some use default geometric patterns, celebrity stock photos, or cartoons). We ignore geometric patterns and rely on heuristics to infer gender from the gender of the person or character depicted in the photo. For example, we infer male from a picture of Kenny McCormick, the South Park character, and female from a picture of Angelina Jolie. Moreover, some of the SO users choose to display pictures of their babies or children, and some display photos of animals, places, or artificial symbols. We chose not to infer gender from such avatars.

We have also observed that people often use the same way to identify themselves (e.g., the same avatar picture, or the same nickname) in other online communities where they are participating (e.g., Github, Twitter, Flickr, or LinkedIn). However, the level of personal information available for a given person in each of these communities may differ. For example, a person's Twitter account may also display her full name, or a person's avatar picture may also be used when she comments on blog posts, where she signs with her full name. Whenever we cannot directly infer gender using the approaches above, we manually investigate the information available from one's participation in other online communities, and try to infer gender from full names, as discussed above.

IV. Pilot StackOverflow Survey

To obtain insights in the demographics of SO, we conducted a pilot survey. We asked the respondents to indicate their SO userid, gender, age, country of birth, country of residence, highest education level obtained and years of professional experience, as well as involvement in open-source and proprietary software development. We obtained 141 responses, including 127 valid ones (e.g., a unique SO userid mapped to an individual). Since the responses were obtained voluntarily, composition of the sample is likely to be affected by a selection bias. However, this data was only used to derive qualitative conclusions.

Our first observation is that the majority of respondents are male: only 12 respondents from 127 have identified themselves as females. It could have been the case that women were less inclined to participate in the survey as the information about age, country of birth or country of residence can be considered private, and they might prefer not to disclose it.

Moreover, we have seen that the respondents are predominantly involved either exclusively in proprietary software (47 respondents) or both in proprietary and open source software (47), while the number of exclusively open source developers was lower (17)[8]. This means that *a priori* one could have expected the share of female SO users to be between 1-5% reported for open source projects [10] and 28% reported for proprietary software [26]. We verify this expectation in Section V. Finally, we have observed that a significant group of respondents (25 out of 127) no longer resides in the countries of their birth due to personal, professional or educational reasons.

V. Results

A. Qualitative analysis

Overall, we observed 2,297 male users, 291 female users, and 1,556 users for which a gender could not be identified, and are not considered as either[9]. These numbers show an overall representation of women at around 7% of the participants, and a vast majority of male users, rejecting the null hypothesis $H_{1,0}$ and accepting the alternative hypothesis $H_{1,1}$. We also found that only a fraction of the selected SO users posed questions, and an even smaller subset answered questions. The boxplots for the distributions of number of questions posed, number of answers given, days engaged and reputation levels achieved, are provided in Figure 1.

B. Hypothesis testing

When analysing the distributions of the numbers of questions (and answers) given, we ignore individuals that did not pose any questions (answers). The differences between genders are visible in the averages and quartiles of the boxplots: statistical tests were therefore used to assess whether such visual differences are also significant.

	n_1	n_2	U	p	$H_{i,0}$
questions	1,237	147	102,673	0.0049	$H_{2,0}$: X
answers	1,004	79	46,175	0.0074	$H_{3,0}$: X
days	1,717	191	193,063	3.3e-05	$H_{4,0}$: X
reputation	229	291	402,246	$<$ 1e-06	$H_{5,0}$: X

TABLE II
MANN-WHITNEY TESTS (STACKOVERFLOW). ✓ - HYPOTHESIS CANNOT BE REJECTED, X - HYPOTHESIS IS REJECTED

[8] The remaining respondents are either not involved in software development at all or they indicated a more elaborate answer than "yes"/"no".

[9] For 616 out of these (or 40%) it is impossible to infer gender, since they have standard SO-assigned user-names (e.g., user1234), no avatar pictures, and no MD5 email hashes in common with other SO users. For most, it was noted a very low reputation on SO, denoting very little activity.

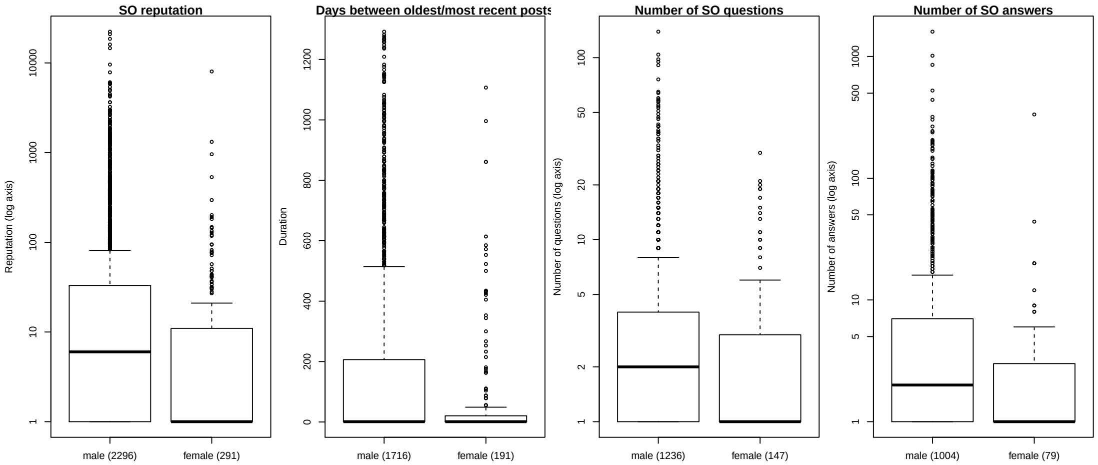

Fig. 1. Boxplots for the studied attributes in the SO sample, divided by gender

The results of the Mann-Whitney tests are reported in Table II. All null hypotheses can be rejected and the corresponding alternative hypotheses should be accepted. The direction of the alternative hypotheses is also significant: men pose more questions, provide more answers, stay involved for longer periods of time and ultimately achieve higher reputation than women.

C. Discussion and Implications

As also noticed by the participants of StackOveflow [36], [37], although the computing field is generally unbalanced towards men, the community around this Q&A site seems to create and maintain higher barriers to entry for women. It does so by designing an approach to collaboration based on earning prizes, achieving higher status and promoting and fostering extremely fast responses by the participants [21], in turn producing more reputation and status. Additionally, the presence of sexism in technology- and computing-oriented communities (as in the programming community around the Linux project) is not only under-estimated, but also frowned upon as a "non-problem" by the male audience [14], [25].

The visible effects of this online behavior are various, but always resulting in declining numbers of women participating in online contribution. Anecdotal evidence [36], [37] suggests:

1) women do not engage with the software development online community;
2) women are turned off by the blatant sexism of participants and leave these communities;
3) women use neuter names or "male profiles" to cope and be accepted by the mostly male audiences.

With specific gender reluctant to participate in online communities, a number of unsolved challenges still persist, from encouraging women to enter the field of technology; to their participation in online communities for expert help and advice; to their sharing of knowledge with the other members. Encouraging women to participate more in sciences and technology has been variously recognized by national and international funding bodies [25], and it is at the heart of the "Science: It's a Girl Thing" initiative recently advertised by the European Commission [1].

VI. Related Work

The first group of related work studies the relation between gender and information technology, or technology in general. Several qualitative studies have focused on the reasons why women are not willing to embark in STEM-related subjects and careers [11], [17], [29]. Various reasons have been given to the under-representation of women in STEM subjects: a general lack of interest in STEM subjects [17], stereotyped thinking by family and teachers [13], lack of role models [23], and most often a combination of various causes together [34]. According to Clance et al [9], the unwillingness observed in minorities group to participate in online communities has created the so-called "imposter syndrome" amongst women: in spite of having good knowledge and being professionally well-settled, women believe they are disqualified or are doing a fraud by fooling others. Nicole Sullivan [32] recommends "Do not feed the Trolls" that can become discouragement amongst the users or members. Apart from the "Science: It's a Girl Thing" initiative launched in June 2012 [1] by the EU commission, there are also some other gender specific online communities who support women in computing and also provide them a private space to take advice from other members in the same field, e.g., the Anita Borg Institute for Women and Technology[10]. Our interest in participation of women and men in online communities can be linked to a recent study of dedication (to programming or to people) [8]. As opposed to thirteen of semi-structured interviews carried out in [8], we consider a much broader group of participants.

The second group of related work focuses on the "use" of computing technology by gender. It has been pointed out that the use of Internet technologies is not as unbalanced as the access to careers and vocational studies [5], and the

[10] http://anitaborg.org

"hacker" culture tends to be male-dominated [33], although the number female hackers is not easy to evaluate [2]. Conversely, a major advocate of the open source phenomenon posit that the hacker culture does not favor a specific gender, rather being asexual when dealing with technology-oriented problems [28]. This has been questioned by recent results showing an "active discrimination" towards women [25].

A number of studies targeted a broader question of differences in the on-line behavior between men and women [16], [27]. Specifically, impact of gender on participation in online communities has been studied in, e.g., communities targeting cancer [12] and travel [40]. Women have been more actively involved in cancer communities then men [12], despite the common observation that computer-mediated communication, in general, is a male-dominant technology and privileges men [30]. In the on-line travel community [40], it has been found that men, holding age and educational level constant, have been community members for longer period of time. These results are concurrent with our observation for StackOverflow.

The third group of related work targets gender resolution, the core step of our empirical analysis. As opposed to using interviews in the aforementioned sociological studies of genders, we have used a heuristics-based name-based gender resolution augmented with manual analysis. Name-based gender resolution has been attempted before (e.g., [15]). However, while [15] reports using name lists for USA only, we employ a much broader search across 30 countries, and use additional heuristics. Alternatively, we could have attempted to recognize genders based on the style of writing [3]. Style-based gender resolution involves counting so called markers that are more frequently used by writers of a certain gender, e.g., pronouns "I", "you" and "she" are significantly more often used by females, while "of"-phrases ("garden of roses") are more typical for male writers [3]. The authors report accuracy of gender resolution to achieve 80% [18]. An obvious advantage of the style-based gender-resolution is its robustness against individuals masquerading as persons of an opposite gender. However, style-based gender-resolution is likely to be affected by the writing style: intuitively, SO questions and answers are neither similar to fiction nor to non-fiction documents (i.e., scientific papers) considered in [3]. Moreover, gender-resolution accuracy will be affected by errors made by non-native speakers. Complementary approaches to gender resolution have been proposed by the image processing community [20]. Ideally, these approaches could have simplified or even replaced the manual avatar analysis. Unfortunately, many avatars cannot be regarded as facial images (symbols, cartoons, body parts). Moreover, application of image processing approaches such as [20] would require a manual preprocessing step involving cropping, resizing and rotation of images.

Finally, Q&A websites, and specifically StackOverflow, are gaining more and more interest from the research community: since 2010, more than twenty research papers were based on the StackOverflow data [39], e.g., [21], [35].

VII. Threats to Validity

The validity of this study is subject to several threats. In the following, threats to *internal validity* (whether confounding factors can influence the findings), *external validity* (whether results can be generalized), and *construct validity* (relationship between theory and observation) are illustrated.

Internal validity – The observation (pilot survey) that a significant number of SO users no longer reside in their birth country can affect the internal validity of the name-based gender resolution. Indeed, names associated with one gender in the birth country may be associated with a different gender in the residence country. Since SO users indicate residence country as their location, this means that the gender-resolution heuristics will make a wrong choice: e.g., an Italy-born male Andrea living in Germany will be identified as female.

External validity – The presented results are only valid for the StackOverflow community. We suspect that other online communities act with similar gender barriers (e.g., gaming communities), while others are more gender- and minorities-friendly (e.g., those related to web technologies).

Construct validity – In addition to threats related to the automatic gender resolution process or the heuristics therein, we note potential human error when inferring gender from an avatar picture, or when deciding whether a certain profile in another data source (e.g., Twitter) belongs to the same SO participant. Another threat to validity refers to SO users purposely using as avatars images of the opposite gender, e.g., male users with erotic stock photos of female models. These cases have been identified and resolved during the manual review.

VIII. Conclusion and Future Work

The issue of gender and STEM-related subjects has been studied for several years, and mostly from the point of view of "why" women do not engage with scientific studies or careers. Lesser attention has so far been given to quantify the phenomenon and representation of women in online communities (as technology-"users"), what are their levels of participation, and whether differences can be detected at the gender level. Only anecdotal evidence has been gathered on how specific communities actively discourage women from participating.

This study quantitatively investigated the participation of women in the StackOverflow Q&A website. The main objective of the study was to add facts to current anecdotal evidence, that suggests that StackOverflow actively discourages the participation of women. In the analysis and attribution of gender to participants of online communities, it was found that a large proportion of SO users are not identifiable. Special tools were developed to infer gender based on name and nationality. However, since the gender inference was partly manual, the SO data was sampled.

It was found that the percentage of women engaged in SO is greatly imbalanced, and men represent the vast majority of contributors. This finding is in line with the recent down-fall in number of graduates in STEM (and computing in particular) subjects. Moreover, women are not only a minority in SO, but their levels of participation are significantly different from men's: men participate more, earn more reputation, and engage in the "game" more than women do.

Future work should expand on the current notion of gender as a binary phenomenon (male/female), an approach that has been already criticised by some of the gender-technology students [8], [38]. Indeed, the conflation of gender and heterosexuality has been observed to complicate social relations in male-dominated domains like computing [31], and lesbian women may feel attracted to software development for the same reasons that heterosexual women may feel disinterested in this field [19]. Therefore, as a possible direction for future work we consider going beyond the gender binary and investigating how does sexual orientation affects individual involvement in online software developers' communities.

References

[1] Science: It's a girl thing! http://science-girl-thing.eu/. Accessed: 30/07/2012.

[2] A. Adam. Hacking into hacking: Gender and the hacker phenomenon. *ACM SIGCAS Computers and Society*, 33(4):3, 2003.

[3] S. Argamon, M. Koppel, J. Fine, and A. Shimoni. Gender, genre, and writing style in formal written texts. *Text*, pages 321–346, 8 2003.

[4] V. R. Basili and D. M. Weiss. A methodology for collecting valid software engineering data. *Software Engineering, IEEE Transactions on*, SE-10(6):728–738, 1984.

[5] B. Bimber. Measuring the gender gap on the internet. *Social Science Quarterly*, 81(3):868–876, 2000.

[6] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan. Mining email social networks. In *MSR*, pages 137–143. ACM, 2006.

[7] J. Blickenstaff. Women and science careers: leaky pipeline or gender filter? *Gender and Education*, 17(4):369–386, 2005.

[8] I. Boivie. Women, men and programming : Knowledge, metaphors and masculinity. In S. Booth, S. Goodman, and G. Kirkup, editors, *Gender Issues in Learning and Working with Information Technology: Social Constructs and Cultural Contexts*, pages 1–24. IGI Global, 2010.

[9] P. Clance. *The impostor phenomenon: overcoming the fear that haunts your success*. Peachtree Publishers, 1985.

[10] P. A. David and J. S. Shapiro. Community-based production of open-source software: What do we know about the developers who participate? *Information Economics and Policy*, 20(4):364–398, 2008.

[11] A. Fisher, J. Margolis, and F. Miller. Undergraduate women in computer science: experience, motivation and culture. *ACM SIGCSE Bulletin*, 29(1):106–110, 1997.

[12] T. Ginossar. Online participation: a content analysis of differences in utilization of two online cancer communities by men and women, patients and family members. *Health Communication*, 23(1):1–12, 2008.

[13] A. Gras-Velazquez, A. Joyce, and M. Debry. Women and ict. *Why are girls*, 2009.

[14] V. Henson. HOWTO Encourage Women in Linux. http://tldp.org/HOWTO/Encourage-Women-Linux-HOWTO/, 2002.

[15] A. Herdağdelen and M. Baroni. Stereotypical gender actions can be extracted from web text. *J. Am. Soc. Inf. Sci. Technol.*, 62(9):1741–1749, Sept. 2011.

[16] S. C. Herring. Gender and democracy in computer-mediated communication. In R. Kling, editor, *Computerization and Controversy: Value Conflicts and Social Choices*, pages 476–489. Academic Press, San Diego, CA, USA, 2nd edition, 1996.

[17] C. Hill, C. Corbett, and A. St Rose. *Why So Few? Women in Science, Technology, Engineering, and Mathematics.* ERIC, 2010.

[18] M. Koppel, S. Argamon, and A. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[19] C. Landström. Queering feminist technology studies. *Feminist Theory*, 8(1):7–26, 2007.

[20] X. Lu, H. Chen, and A. Jain. Multimodal facial gender and ethnicity identification. In D. Zhang and A. Jain, editors, *Advances in Biometrics*, volume 3832 of *Lecture Notes in Computer Science*, pages 554–561. Springer Berlin / Heidelberg, 2005.

[21] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann. Design lessons from the fastest q&a site in the west. In *Human factors in computing systems*, pages 2857–2866. ACM, 2011.

[22] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[23] J. Margolis and A. Fisher. *Unlocking the clubhouse: Women in computing*. The MIT Press, 2003.

[24] I. Miliszewska, G. Barker, F. Henderson, and E. Sztendur. The issue of gender equity in computer science- what students say. *Journal of Information Technology Education*, 5(1):107–120, 2006.

[25] D. Nafus, J. Leach, and B. Krieger. FLOSSPOLS Deliverable D 16 Gender: Integrated Report of Findings. http://www.flosspols.org/deliverables/D16HTML/FLOSSPOLS-D16-Gender_Integrated_Report_of_Findings.htm, 2006.

[26] National Science Foundation. Women, Minorities, and Persons with Disabilities in Science and Engineering. http://www.nsf.gov/statistics/wmpd/pdf/nsf04317.pdf, 2004. NSF 04-317.

[27] M. R. Parks. Making friends in cyberspace. *J. Computer-Mediated Communication*, 1(4), 1996.

[28] E. Raymond. The cathedral and the bazaar: Musings on linux and open source by accidental revolutionary revised edition. *Cambridge, UK, O'Reily*, 1999.

[29] E. Roberts, M. Kassianidou, and L. Irani. Encouraging women in computer science. *ACM SIGCSE Bulletin*, 34(2):84–88, 2002.

[30] L. H. Shaw and L. M. Gant. Users divided? exploring the gender gap in internet use. *CyberPsychology & Behavior*, 5(6):517–527, 2002.

[31] L. Stepulevage. Gender/Technology relations: complicating the gender binary. *Gender and Education*, 13(3):325–338, 2001.

[32] N. Sullivan. Don't feed the trolls. http://www.youtube.com/watch?v=ulNSlES1Fds, 2012.

[33] P. Taylor. *Hackers: crime in the digital sublime*. Psychology Press, 1999.

[34] E. Trauth, J. Quesenberry, and A. Morgan. Understanding the under representation of women in it: Toward a theory of individual differences. In *Proceedings of the 2004 SIGMIS conference on Computer personnel research: Careers, culture, and ethics in a networked environment*, pages 114–119. ACM, 2004.

[35] C. Treude, O. Barzilay, and M.-A. D. Storey. How do programmers ask and answer questions on the web? In R. N. Taylor, H. Gall, and N. Medvidovic, editors, *ICSE*, pages 804–807. ACM, 2011.

[36] VA. Never see any women answering questions? http://meta.stackoverflow.com/q/34070/185480, 2011.

[37] VA. What can Stack Overflow do to persuade female programmers to participate more? http://meta.stackoverflow.com/q/30411/185480, 2011.

[38] A. Van Lenning. The body as crowbar: Transcending or stretching sex? *Feminist Theory*, 5(1):25–47, 2004.

[39] B. Vasilescu. Academic papers using Stack Overflow data. http://meta.stackoverflow.com/q/134495/185480, 2012.

[40] Y. Wang and D. R. Fesenmaier. Modeling participation in an online travel community. *Journal of Travel Research*, 42(3):261–270, 2004.

[41] M. A. Whitecraft and W. M. Williams. Why arent more women in computer science? In A. Oram and G. Wilson, editors, *Making Software: What Really Works, and Why We Believe It*, pages 221–238. O'Reilly Media, Inc., 2010.

[42] R. T. A. Wood and M. D. Griffiths. Why swedish people play online poker and factors that can increase or decrease trust in poker web sites: A qualitative investigation. *Journal of Gambling Issues*, 21:80–97, 2008.