BACHELOR THESIS

From Asking to Answering: Social and Structural Aspects of Contributing to Stack Overflow

submitted by

TIMUR BACHSCHI

Submitted to the

Chair of Computational Social Sciences and Humanities

within the

Faculty of Mathematics, Computer Science and Natural Sciences

at RWTH Aachen University

March 3, 2020

First reader:

Prof. Dr. Markus Strohmaier

Second reader:

Prof. Dr. Ulrik Schroeder

Advisor:

Dr. Johannes Wachs

Zentrales Prüfungsamt/Central Examination Office

Eidesstattliche Versicherung
Statutory Declaration in Lieu of an Oath

Name, Vorname/Last Name, First Name

Matrikelnummer (freiwillige Angabe)
Matriculation No. (optional)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/ Masterarbeit* mit dem Titel
I hereby declare in lieu of an oath that I have completed the present paper/Bachelor thesis/Master thesis* entitled

selbstständig und ohne unzulässige fremde Hilfe (insbes. akademisches Ghostwriting) erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.
independently and without illegitimate assistance from third parties (such as academic ghostwriters). I have used no other than the specified sources and aids. In case that the thesis is additionally submitted in an electronic format, I declare that the written and electronic versions are fully identical. The thesis has not been submitted to any examination body in this, or similar, form.

Ort, Datum/City, Date

Unterschrift/Signature

*Nichtzutreffendes bitte streichen

*Please delete as appropriate

Belehrung:
Official Notification:

§ 156 StGB: Falsche Versicherung an Eides Statt
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.
Para. 156 StGB (German Criminal Code): False Statutory Declarations
Whoever before a public authority competent to administer statutory declarations falsely makes such a declaration or falsely testifies while referring to such a declaration shall be liable to imprisonment not exceeding three years or a fine.
§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt
(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.
(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.
Para. 161 StGB (German Criminal Code): False Statutory Declarations Due to Negligence
(1) If a person commits one of the offences listed in sections 154 through 156 negligently the penalty shall be imprisonment not exceeding one year or a fine.
(2) The offender shall be exempt from liability if he or she corrects their false testimony in time. The provisions of section 158 (2) and (3) shall apply accordingly.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:
I have read and understood the above official notification:

Ort, Datum/City, Date

Unterschrift/Signature

Contents

1 Introduction and Overview

The Q&A platform Stack Overflow is a primary source of information for people who write code. The site typically occupies top search engine responses to queries about programming [42]. Acting as a collectivized knowledge hub, it not only offers answers to specific questions but also provides people with resources to learn new concepts or even programming as a whole. In this way, Stack Overflow provides an essential piece of infrastructure to the open-source software (OSS) ecosystem [14]. Open-source developers can turn to Stack Overflow to speed up their work and also to interact with their users [40].

Open source software itself faces the interrelated issues of sustainability and lack of diversity. Research indicates that while the number of people being able to code and the usage of open-source code have increased, the frequency of new contributions to these projects has not kept up. Many such projects have become the unpaid full-time job of just a handful of contributors [14]. While searching for reasons for this development, some researchers have focused on finding possible barriers for new users to contribute [39]. Finding and weakening these barriers in the pipeline from novice coder to experienced open-source contributor is essential to building a more sustainable and diverse ecosystem. In a similar fashion, a decrease in new contributors is also observable on Stack Overflow. Figure 1 shows the share of users on Stack Overflow who contribute to the knowledge base with answers decreasing over the years.

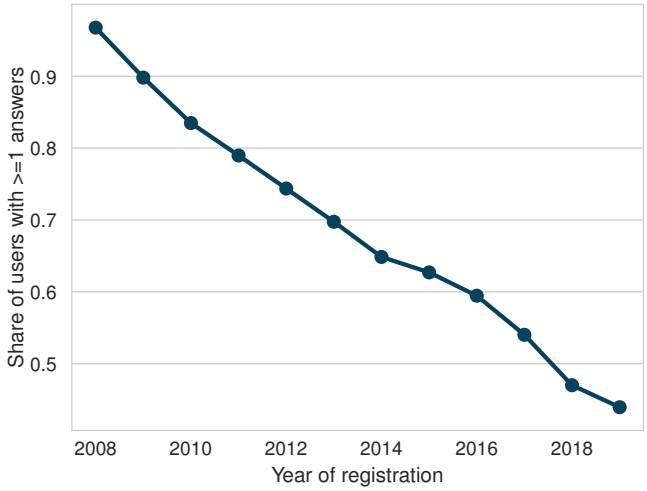


Figure 1: *Share of users with atleast one answer post by the year of user registration. Over the years, the share of users interacting with Stack Overflow by posting answers has decreased significantly.*

This trend suggests that fewer people joining the platform are becoming active contributors. This presents problems both for the open-source world in general and for Stack Overflow. Previous research has documented a link between providing answers on Stack Overflow and contributing to projects on GitHub [40]: it is possible that answering questions on Stack Overflow serves as a stepping-stone to making contributions to open-source software libraries. Useful documentation of software via contributions on Stack Overflow may also relieve pressure on core developers

by answering questions that may otherwise be asked multiple times via email or mailing lists. In this way, rich Stack Overflow activity around an open source library generates an evolving, crowdsourced Q&A. The existence of such a knowledge base for a library may even facilitate onboarding to the library's development team itself [32].

This perspective - that Stack Overflow serves as a knowledge base for users of software - highlights a second potential negative consequence of the trend observed in Figure 1 The most active contributors of knowledge to Stack Overflow tend to have registered early on. Such core contributor populations are typically even less representative of society as a whole than the software engineering community in general [14]. If certain subpopulations of the user base such as women or people from developing countries are less likely to contribute answers, the knowledge base will, over time, neglect their perspectives. This may lead to a vicious feedback cycle - if the knowledge sources reflect a skewed perspective, members of excluded communities will have a harder time joining later [8]. Similar effects are observable on Wikipedia, another prominent example of a decentralized knowledge database [41, 31].

More practically, answering questions on Stack Overflow can build confidence and serve as an intermediate step between learning and contributing more directly to open source projects. Therefore, it is of interest to detect barriers that hinder Stack Overflow users in transitioning from asking questions to providing answers on the platform. In this thesis we mine the Stack Overflow data dump and model the likelihood a user will make this shift using personal and community-level features. We find that users putting additional effort into maintaining their Stack Overflow profile show faster and more likely questioning-to-answering transitions, which may relate to the self-promotional aspect of the platform. Besides that, we observe barriers for groups underrepresented on the platform like women or users from the global south. Regarding transition, the communities a user engages with change behavior with more negative and hostile ones having a deterrent effect. Zooming in on the *Swift* programming language, established in 2014, we find that previous experience on the platform in non-Swift communities makes transition faster and more likely. The type of previous experience plays a role in that previous interaction in communities closely related to Swift relates to faster transitioning.

1.1 Research Question

The primary goal of this thesis is to understand the social and behavioral features of Stack Overflow users that predict when they will transition from asking questions to answering them. We accomplish this goal by modeling this transition utilizing data from all posts made on Stack Overflow since its creation in 2008. The primary research question can be stated as follows:

Primary Research Question: What user activities, social attributes, and community factors are significant predictors of the number of questions until transition to answering questions on Stack Overflow?

To answer this research question we examine various features of the behavioral traces of Stack Overflow users and how those relate to the likelihood they transition. We categorize those features into three groups: individual, tag-level and site-level. We then test the size and statistical significance of these features in models predicting how quickly users post their first answer, and if they contribute an answer at all. We find that features at all three levels have a statistically significant relationship with how long it takes for a user to make an answering contribution. The findings suggest that barriers to becoming a contributor are multifaceted.

1.2 Related Work

Several previous works analyze posting behavior and user careers on Stack Overflow. Slag et al. look at users who only post once on Stack Overflow and then drop out, analyzing possible reasons for this behavior [33]. They found out that posts by the so-called "one-day flies" are deleted more often and are more likely to receive no answer, which may cause the disengagement with the website. In this thesis, we are exclusively going to focus on users with a certain level of engagement on Stack Overflow and how the type of their interaction changes and why. The so-called "one-day flies" are ignored in this thesis. One could also inspect the content of posts to find out possible transition barriers on a linguistic basis. For example, Danescu-Niculescu-Mizil et al. provide a model for analyzing politeness levels of posts on Stack Exchange, which could be used for such an analysis [12]. To tackle the negativity newcomers on Stack Overflow receive when asking their initial questions, Ford et al. deployed a mentorship program to help new users in the formulation of their questions in an on-site *Help Room* [18]. The surveyed users that participated in this program agreed to feel more comfortable posting on Stack Overflow after participation. Finally, recent work has shown that women are signficantly more likely to post questions and less likely to post answers than men [24]. This suggests just one way social factors relate to observed behavior on Stack Overflow.

2 Data and Features

To answer our research question we use data about Stack Overflow and its users. We obtain this data from the Stack Exchange Data Dump[1], which contains all the user-contributed content from the Stack Exchange network, including Stack Overflow. Our primary analysis uses the data dump from June 2019, though we also consider previous data dumps in order to analyze the evolution of tag-communities later in the thesis. We now proceed to describe the data used in the thesis.

2.1 Data

The Stack Exchange data dump provides multiple XML files containing various information regarding Stack Overflow. The ones of interest to us are the data about posts, users and votes. The scripts used to process and analyze this data are available on GitLab[2]. Table 1 presents a selection

[1] https://archive.org/details/stackexchange

[2] https://git.rwth-aachen.de/timurbachschi/from-asking-to-answering-so

of descriptive statistics for the data. Of the 45 million posts, around 27 million are answers and 17 million are questions, meaning there are more answer than question posts. The approximately 100,000 posts which are neither questions nor answers, like moderator nominations, are excluded in our analysis.

Table 1: *Statistics for the Stack Overflow Dataset.*

#Posts	44,945,355
#Answers	27,107,580
#Questions	17,738,809
#Other posts	99,066
#Users with 5+ posts	1,188,419

Table 2 shows the attributes of the post, user and vote data that we will use later on. A post consists of a unique ID, a type, a creation date, the number of views, the ID of the creator and the set of all tags. If the post is an answer, it additionally has the ID of the question that has been answered (also called the parent post). A user has a unique ID, an attribute for the creation date of the account and a username. Besides that, three attributes may or may not be specified by the user:

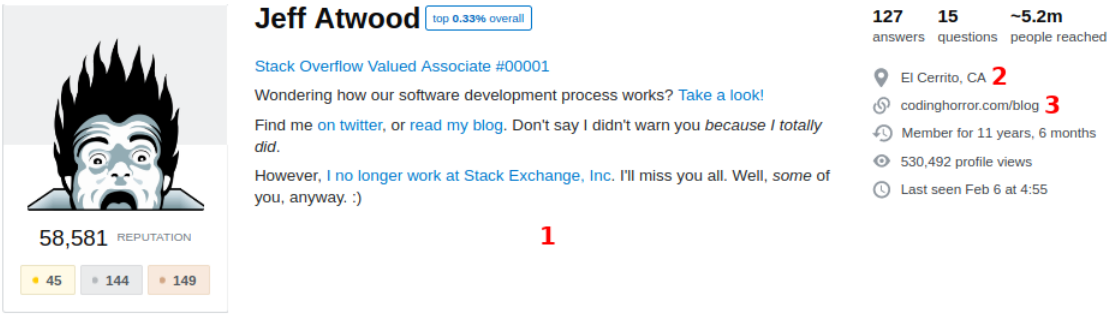


Figure 2: *Stack Overflow profile page containing an About Me section (1), a geographic location (2) and a website URL (3).*

location information, a link to a website (possibly their own) and an About Me section, where a user can write a text describing themselves. Figure 2 shows how these three attributes are displayed on the user's profile page. Besides a unique ID, a vote has an attribute for the ID of the post the vote is made on, the type i.e. upvote or downvote and the creation date of the vote.

In this thesis, we will only look at the number of questions until transition for a subset of all users with at least five posts on the website. We call these *active users*. We also exclude deleted posts. Before going further, we need to distinguish between two types of transition. We define the *general transition* by looking at the number of posts before the user's first answer (including the transition post). The definition for *subcommunity transition* looks similar, but for that only posts from one specific subcommunity/tag are considered. An alternative way of defining when the questioning-to-answering transition happens is to look at the actual time until that shift happens. For simplicity, we prefer using the post count as the way to calculate when the shift happens. Calculating the Spearman correlation coefficient between the question count and time

Table 2: *Description of the data obtained for posts, users and votes from the Stack Overflow data dump.*

Posts	
Id	Identifier for post
PostTypeId	1: Question 2: Answer
ParentId	Id of parent post if post is an answer
CreationDate	Creation date of post (UTC)
ViewCount	Number of views
OwnerUserId	Id of the postcreator user
Tags	Post's tags

Users	
Id	Identifier for user
CreationDate	Registration date of user (UTC)
DisplayName	Name of user
WebsiteUrl	Website URL linked on user's profile
Location	Location info provided on user's profile
AboutMe	Text of AboutMe section of profile page

Votes	
Id	Identifier for vote
PostId	Id of the post for which the vote is cast
VoteTypeId	2: Upvote 3: Downvote
CreationDate	Date of vote casting

until transition results in a high value of $\rho = 0.91$, showing that we can use the simpler measure without obtaining significantly different results. We acknowledge that an analysis focusing on the temporal aspects more directly may be of interest for future work.

Before utilizing the data dump, it needs to be checked for correctness. To do so, we pick a sample of 200 users and compare the data provided and the general transition times calculated utilizing it with their actual profiles on Stack Overflow. The post count until transition and creation date are equivalent on the website and in the data for all 200 users. For 99% of them, the name and About Me section are identical, while the website URL and location are identical in 98% and 97.5% of users respectively. These values highlight the accuracy of our data, especially because we assume some of the discrepancies to stem from the time difference between the data set and the website information.

Our definition of questioning-to-answering transition results in four specific types of users that may appear:

– Users with x question posts and an answer post afterward. Those users transition after x + 1 question posts.

– Users with only answer posts. As they do not post any questions, those *immediate contributors* transition on their first post.

– Users with only question posts who have not stopped engaging on Stack Overflow. Those users may transition at some point later on.

– Users with only question posts who have stopped engaging on the site. Those users will not transition at all.

The first two types of user can be assigned a number of posts until transition without any problem. With the other two, we cannot define such a number. This factor needs to be kept in mind when analyzing user transition. Figure 3 shows the distribution of posts until transition for all users that do transition after one to ten questions. Most of these users appear to be immediate contributors. With increasing number of posts, the share of users that do take that long to transition

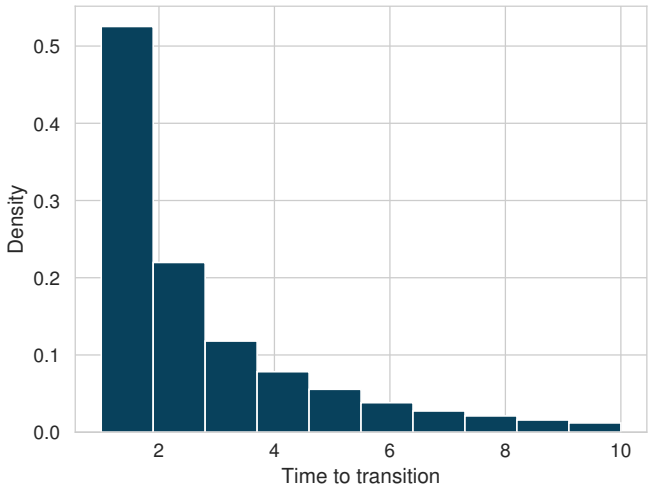


Figure 3: *Shares of users transitioning after 1 to 10 posts. The mode lies at 1. From there, the counts decrease for every increase in the questions count until transition.*

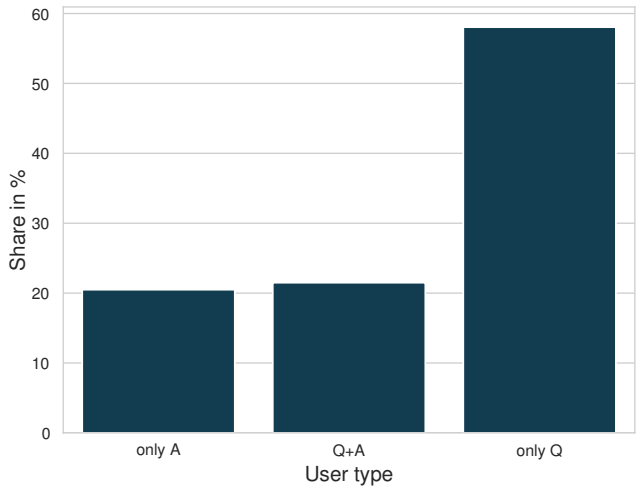


Figure 4: *Share of active users who only post questions (Q), only answers (A) or both (Q&A) on Stack Overflow as a whole, with the latter being the most substantial fraction.*

decreases. The mean value is 3.85 posts until transition. Figure 4 shows the distribution of users who only post questions, only answers, or both. 57% never complete the transition just posting questions, emphasizing the need for removing barriers and making answering on Stack Overflow more attractive. To entice people to provide knowledge for the platform, the causes for the lack of transition for those users are crucial to study.

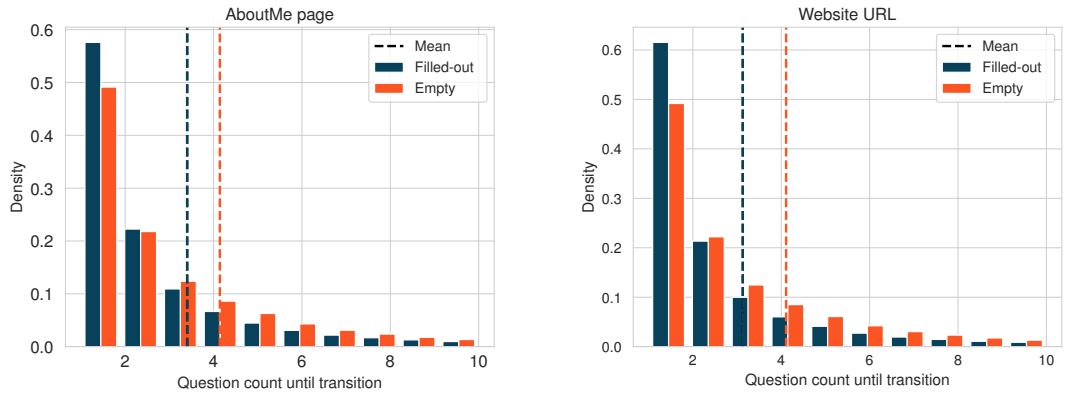
2.2 Individual Features

Individual features describe attributes and characteristics specific to a particular user. We consider the user's profile page, their gender (inferred from their username and location), location, posting time and previous experience. We now discuss how we operationalize these features and explain in more detail how they may relate to a user's likelihood and speed of contributing.

2.2.1 Profile Page

Each user on Stack Overflow has a profile page where their location and web presence (Website/Twitter/Github) can be made public. Additionally, the profile contains a section that we call the *About Me* section, where a user can include further information about themselves in text form. The profile also reveals the posting history. Making it possible to share such information supports social transparence, building trust between the community members [15]. Users providing such data may be seen as more reliable or reputable, similar to what has been observed on the German Wikipedia. On that platform, the high reputation of users, assessed from their profile, increases the probability of articles with contribution by these users to become "featured" [38]. The way the profile and online activity in social programming communities affect public image is well known, even by the individuals themselves. Dabbish et al. found out, while analyzing Github's community, that members were actively managing their public image on the website [11]. Besides profile information, a user on Stack Overflow can also enter job preferences, which will be shown to companies if the individual expresses interest in working for them. This incentivizes users looking for job applications to invest time in constructing their profile. By looking at shifts after employment changes, Xu et al. have found that people signal their competence through their Stack Overflow contribution to appeal to the job market [43]. They have observed a decrease in reputation-generating activity by 23.7% after finding a new job for users from the US and Canada. Of this decrease, 12.5-16.5% could be assigned to a drop in career concerns. Therefore, users who fill out their profile may deliberately try to tie their Stack Overflow profile and identity to their real identity and take the website more seriously. They do so to bolster their reputation in the coding world. Accordingly, we hypothesize that those users will transition between questioning and answering after fewer posts and are also far more likely to do so at all than users who do not place so much effort into maintaining their public image on the website.

As stated before, each individual on Stack Overflow has a profile page containing an "About Me" section. A user does not need to write something in this section, though; many of them do not do so. Our dataset contains the text provided in the description field of the About Me section. The data field is empty, if there is no unique text instead of the standard text of "There appears to be an air of mystery about this user". Figure 5 shows two distributions. The first one presents the post count until transition for users with filled-out About Me pages compared tp those with empty ones. The second one shows posts until transition for users linking to a website in the designated space on their profile. We exclude non-transitioning individuals. The distributions show that users with filled-out About Me sections or a website link on their profile transition after fewer posts on



(a) (b)

Figure 5: *Distributions of question count until transition (capped at 10) of users that do have a filled-out About Me section (a) / website URL (b) vs. those who do not. The share of users that transitioned previously with 1 or 2 questions beforehand is larger, while being lower in all other cases, for users that have empty About Me/website URL sections compared to those with filled-out ones. The latter group also has the higher mean number of posts until transition. The median count of questions until transition for users with website URLs is 1 and 2 for every other group.*

average with many more immediate contributors. The share of non-transitioners is around 11.8% for users with filled-out About Me sections and 40.4% for those with empty ones.

2.2.2 Gender

There are gender differences in regard to Stack Overflow participation. Only 7.9% of all Stack Overflow contributors declare to be women [35]. Several potential barriers exist that result in this low number, for example, significant fear of negative feedback [17]. Women also have lower reputation scores and give fewer answers than men [24]. Providing fewer answers also relates to the often observable understatement of their own technical abilities [28]. On top of that, it has been discovered that women are less likely to fill out their personal profiles [4]. All these aspects may occupy a significant role in completing the questioning-to-answering transition. We hypothesize that these already known barriers play a role in transitioning in that women are less likely to complete the shift and take more posts to do so. Stack Overflow does not provide a way to specify a user's gender or displaying that on their profile page. Therefore, we need to infer gender differently. First we must acknowledge that we, like previous quantitative studies of gender gaps on Stack Overflow, make the simplifying assumption that gender is binary and that binary gender can accurately be inferred from screen names.

Many usernames are nicknames from which we cannot infer gender, but there are also many usernames containing either their first-name or last-name or both. Therefore, we run the Python library *Gender-Guesser*[3] on all nicknames and, if possible, on both strings separated by the first occurring space. We do that to catch the cases where the username represents a combination of

[3] https://github.com/lead-ratings/gender-guesser

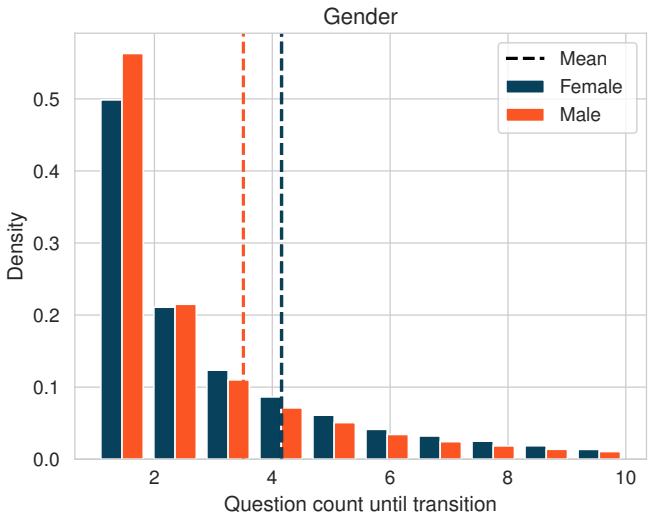


Figure 6: *Distributions of question count until transition (capped at 10) of female vs. male users. The share of users that transitioned previously with 1 and 2 questions beforehand is larger, while being lower for every other number of posts until transition, for male users compared to female ones. The The latter group has the larger mean post count until transition of both. For both groups, the median post count is two.*

first-name and last-name. The library uses a lookup table to match names to gender. Using this technique results in 23.7% of all individuals getting classified as either male or female by the algorithm. For analyzing the difference in transition concerning gender, we solely look at the cases where a user is categorized as male or female. Figure 6 shows the distribution of posts until transition concerning this category for transitioning users. Male-classified users appear to make the shift after fewer posts on average compared to female-classified ones. Additionally, the share of immediate contributors is higher for male-classified users. The share of non-transitioning ones is 26.7% for male-classified and 44.5% for female-classified users.

2.2.3 Geography

While the users of Stack Overflow come from various locations, Oliveira et al. show that the percentage of contributing individuals vary from country to country [27]. They highlight two factors having an effect on this divide. The first one, especially applicable for users from outside the anglosphere, is the language barrier faced on the website. English is the default language of Stack Overflow [36], causing barriers for non-native English speakers using the platform [20]. The Stack Overflow developer survey confirms this, as it shows that English speaking skills are a crucial factor in the user's decision to participate on the website, as shown in Figure 7 Language barriers have also caused a division in the SO community through the creation of localized variations of the website like a Russian one, for example. The second factor regarding the varying contribution count are differences in national cultures. By looking at users from the US, the UK, China and India with similar job roles and employers, Yang et al. have discovered differences regarding questioning

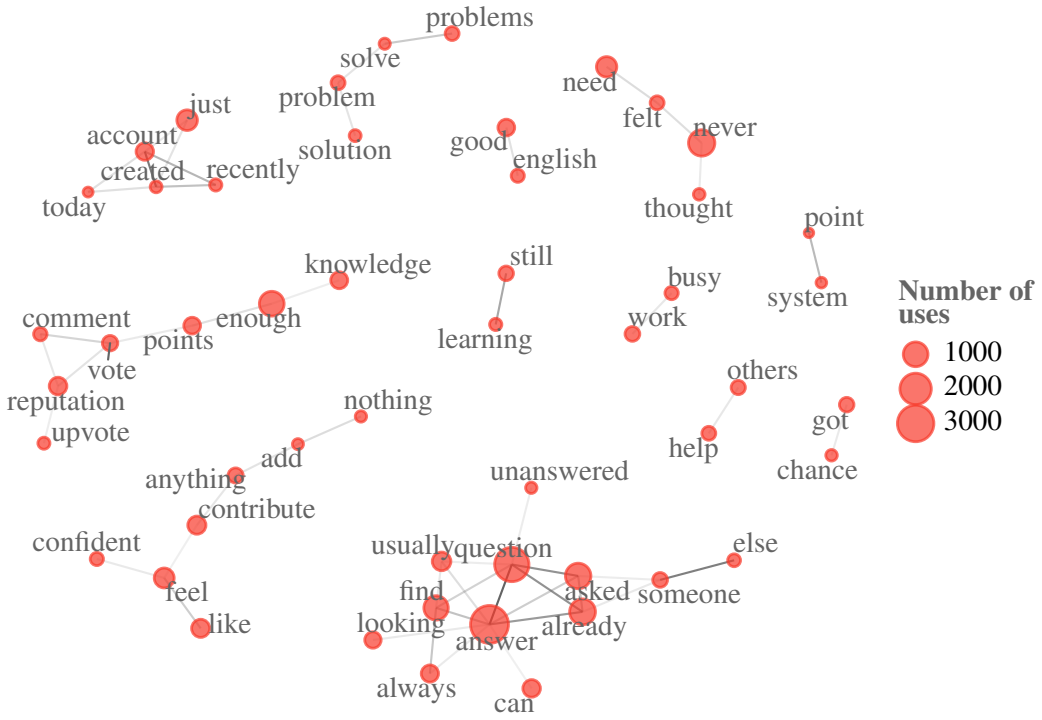


Figure 7: *From Stack Overflow developer survey [35]: Network shows the most correlated words among 10,692 responses to the question for reasons not to participate in Q&A; words with bigger bubbles are used more often, and heavier connections indicate terms more correlated with each other. We see a relationship between "good" and "english".*

and answering on online Q&A platforms across the four countries [44]. More generally, they show a difference between the western and eastern ones. They argue that this distinction relates to difference between western individualistic and eastern collectivistic cultures. Besides cultural and linguistic reasons for varying levels of contribution, there are also political and economical ones at play. Countries have different infrastructure, policies and education levels. Such differences have resulted in a worldwide digital divide, where countries from the global south may not have the same growth and access to Internet technologies as those from the global north [9]. All these aspects make analyzing how the geographic location and transition of a user correlate worthwhile. We hypothesize that users from the global south transition slower and less likely than those from the global north, primarily because of the language barrier.

To assess the effect of location on transition, we can consider the locations given on the user profiles. In the dataset, this information is provided as a string individually set by the user. As there is no set format, it may be given as a city name, a country or both. Additionally, there are some joke locations ("In another galaxy") and cases without any location set at all. To extract a country from those strings, we use the Python library *Geotext*[4]. It allows us to write a program that accepts a string as input and returns every country mentioned in it. There are some cases where the library's function fails the task, though. Cases of countries mentioned in languages besides English

[4] https://github.com/elyase/geotext

are not correctly classified. For example, a mention of Germany as “Deutschland” is not correctly identified by Geotext.

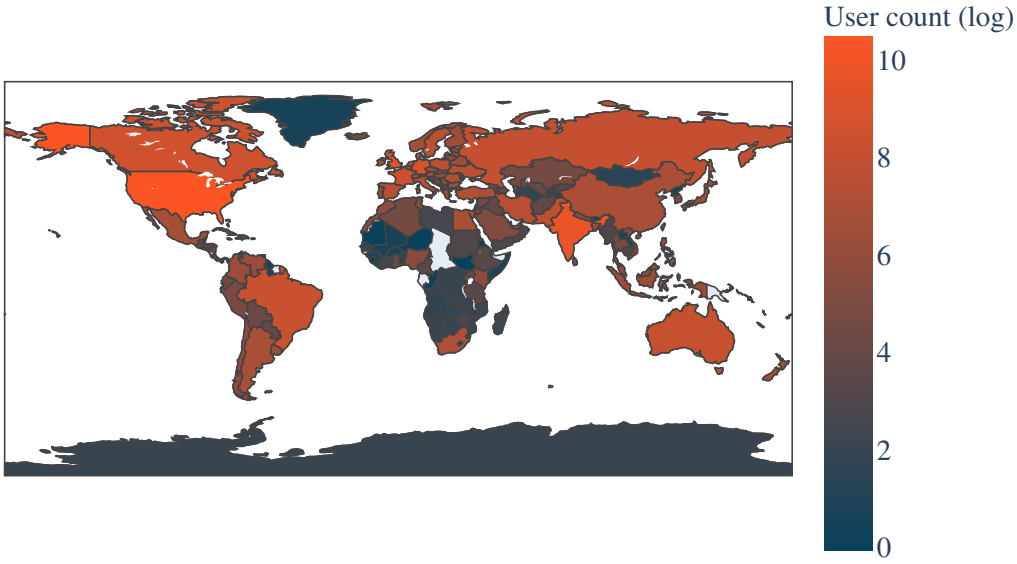


Figure 8: *Visualization of counts of registered users by the location obtained from their profile information. The vast majority of users appear to be from the US. Besides the US, many users seem to be from India, the UK, Germany, Canada, and Russia.*

We also have no direct way of inferring the location of users with this information from their Stack Overflow profiles. This counts for nearly 53% of all users. Figure 8 visualizes the share of country mentions for the users for which we can obtain some geographic data. It shows that most users are from the US, India and Europe. The previously mentioned Stack Overflow developer survey also contains data about location. The survey shows most users are from the US (23.6%) , followed by India (10.2%) , Germany (6.6%) , and the UK (6.5%) [35]. This is similar to our own classification results. To summarize the locations, we use the *global north/south divide*. To get the exact classification for each country, we use the list provided by the Wikimedia Foundation[5]. Figure 9 shows the distribution of question count until transition for users from the global north compared to those from the global south excluding non-transition users. The average post count until transition for users from the global north is slightly higher than the average for global south users, with the former group having the larger share of immediate contributors. The share of non-transitioning users is around 15.9% for global north users and 18.8% for global south users.

[5] https://meta.wikimedia.org/wiki/List_of_countries_by_regional_classification

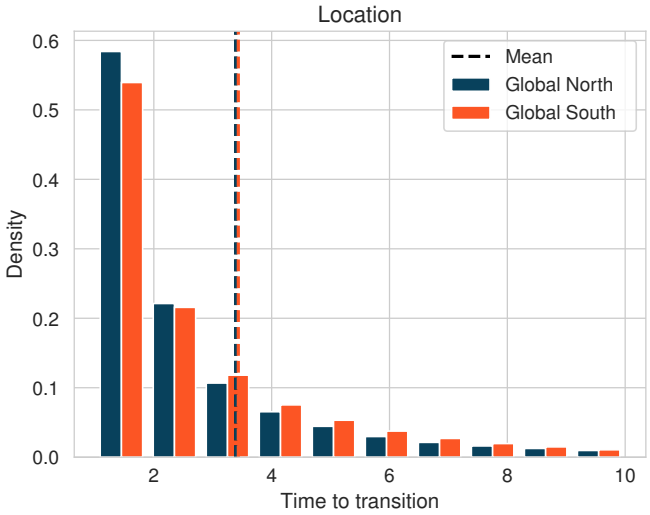


Figure 9: *Distributions of question count until transition (capped at 10) of users from the global north vs. those from the global south. The share of users that transitioned previously with 1 and 2 question posts beforehand is larger, while being lower for every other post count until transition, for global north users compared to those from the global south. The latter group has the (slightly) larger mean post count of both. For both groups, the median post count is 2.*

2.2.4 Weekend Posters

By analyzing questions on the website, Allamanis et al. discovered a difference in participation on weekends compared to weekdays [3]. Weekends appear not only to be less "busy" but also feature other popular tags than weekdays. In that paper, the hypothesis is that this might have something to do with hobbyist coders. Those are more likely post on weekends compared to weekdays, where people use Stack Overflow while at their job. Of all users on Stack Overflow, 80.2% state that they are writing code as a hobby outside of work [35]. We assume a difference in transition for different tags, but besides that, the type of user and their coding ability may equally affect transition. We aim to use the time period a user mostly posts in to approximate whether they are a hobbyist or professional coder. Claes et al. have looked at multiple Git and Mercurial repositories to analyze working times of programmers [10]. They have found a correlation between the amount of work done on weekends and on weekdays. Users that work more during the week also tend to work more on weekends. In addition, the findings by Bosu et al. show that questions on weekends are generally more likely to be answered even though it might take longer [6]. Combining those results, we argue that weekends feature a larger share of users putting significant time and effort into their work, that are more likely to answer a question compared to weekdays. Therefore, we hypothesize that weekend users are more likely to transition and do so faster than weekday ones. Therefore, it makes sense to consider this attribute in our transition models.

We define a user as a weekender if at least 2/7 of their posts are made on weekends. This is a conservative definition of someone who codes primarily on the weekend against the null assump-

tion that posting on Stack Overflow is equally likely any day of the week. Our dataset gives us the exact creation time of a post; all converted to UTC. Utilizing those times to find out whether a post is made on the weekend overlooks the time zone a user is actually in. This may result in some incorrect classifications, a user posting on Monday, 1:00 CEST would be classified as a weekend poster for example. We assume that these errors are uncommon and that borderline cases are not substantively consequential to our analysis (as a post made on Sunday, 23:00 or Monday, 1:00 can be classified as a weekend post either way without losing too much of the reasoning behind the label). Figure 10 shows the distribution of post counts until transition for those two types

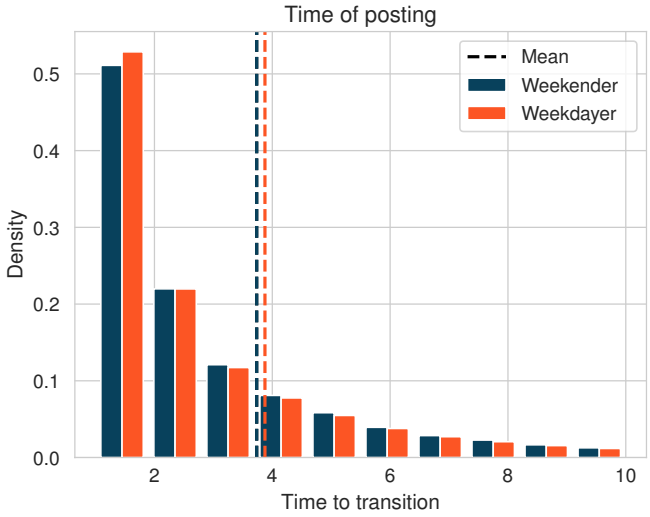


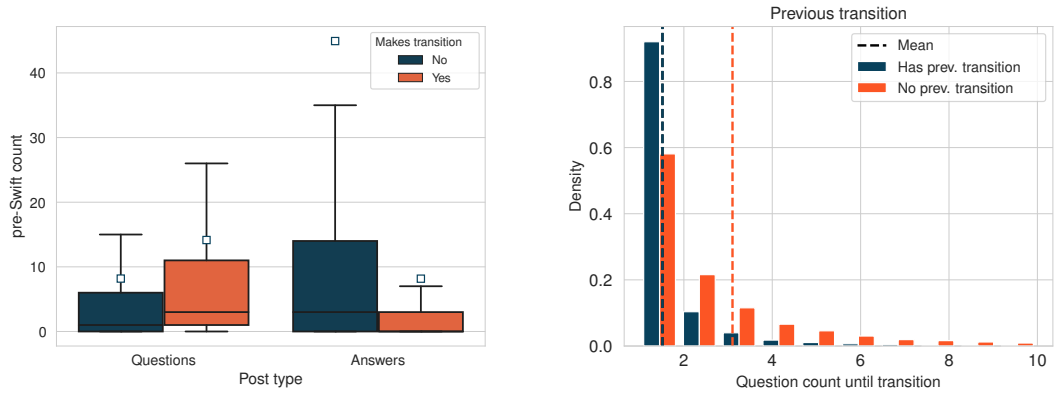
Figure 10: *Distributions of question count until transition (capped at 10) of users that do post on weekends vs. those who post on weekdays. The share of users that transitioned previously with 1 question beforehand is larger, while being lower for all other post counts until transition, for weekday users compared to weekenders. The latter group has the lower mean post count until transition of both. For both groups, the median post count is 2.*

of users excluding non-transition users. The share of non-transitioning users is around 36.8% for weekenders and 30.1% for weekdayers. We see that the average pre-transition post count is slightly lower for weekday compared to weekend users, even though the share of immediate contributors is larger for weekday users.

2.2.5 Previous Experience

There are different types of previous knowledge a user on Stack Overflow might have. For example, when creating an account, a user has some level of experience regarding programming. A new user might also be new to programming as a whole, using Stack Overflow to actively acquire a new technology. On the other hand, a user might also be an expert coder, asking more specific questions. Another type of previous experience which we want to focus on is a previous background on the website itself. For example, we argue that there is a difference between a user stepping into a new SO subcommunity after already posting in different ones compared to someone who is joining a

subcommunity simultaneously with joining Stack Overflow. In their study, Steinmacher et al. [39] have found entry barriers to open source projects. One of these barriers is previous knowledge or the lack thereof. It is more challenging to integrate a person with no prior experience about parts of the project compared to someone who already knows some facets. We can change this barrier to match the Stack Overflow context by looking at tags. When creating a new question post, a user is asked to input at least one tag that relates to their question. The questions page, which is the most visited page on the whole website[6] does not only show all of them but also highlights posts where one of them matches one of the tags set as a technology of interest at the account creation step. We view tags on Stack Overflow as subcommunities. We hypothesize that an active user regarding one subcommunity is going to encounter fewer problems integrating into a different one, especially if they have previously transitioned in another tag. We say this because a user already active on the website, even though in a different subcommunity, still has collected experience about the website. More specifically, one could ask if the type of previous community affects transition similar to how the previous experience of parts of the projects influences the integration of open-source developers. For example, transition for Swift could be slower on average for users who posted in the Objective-C tag before, because of the shared context of those languages [19]. To assess the effect of previous SO experience, we choose a subcommunity and look at the previous



(a) (b)

Figure 11: *(a) Distributions of question and answer counts of users that do transition vs, those who do not. Transitioners appear to have a higher question count on average, while having a lower answer count compared to non-transitioners. (b) Distributions of question count until Swift transition (capped at 10) of users that do have a previous transition in a non-Swift tag vs. those who do not. The share of users that transitioned previously 1 post beforehand is close to 100%. Their mean is also lower than that of users without previous experience. For the previous transitioned group, the median post count is 1, for the other group it is 2.*

experience of users entering that community. For that, we look at the Swift community. Swift is a "robust and intuitive" general-purpose programming language developed by Apple[7]. We pick Swift for our task because of multiple reasons. The first one is that Swift, with its release in June

[6] https://stackoverflow.blog/2017/03/09/anyone-actually-visit-stack-overflows-home-page/
[7] https://www.apple.com/swift/

2014, is younger than Stack Overflow itself. This creates a situation where a significant number of people already have collected experience in other tags, Objective-C in particular. More precisely, Objective-C is the second most frequent tag a user posts in pre-Swift with around 50% of all Swift users having posted in this tag beforehand. Additionaly, the previously experienced users do not merely stay in their old tag but migrate to Swift as seen by the fast growth of the language popularity after it's release [29]. Another reason for choosing Swift is the already mentioned connection to other tags such as Objective-C, providing an opportunity to analyze the relationship between those communities.

Figure 11 depicts multiple distributions related to pre-transition experience. The first one (a) shows the distributions of question and answer counts before the first post in the tag Swift for users posting in Swift at some point in time. By comparing users with Swift transitions and those without, we observe that transitioning users have a higher question count, but a lower answer count pre-Swift. The second one (b) shows the distribution of question counts until Swift transition for users that have a previous made the shift and those that did not. On average, previously transitioned users transition in Swift after significantly fewer posts than users without the previous shift. Former group also has a larger share of immediate contributors. The share of users who do no transition is 32.0% for users with any previous transition and 69.6% for users without.

2.3 Tag-level Features

The experience of a user on Stack Overflow is a social one. A post can get commented on; other users can answer a question. The way those interactions play out may affect transition. As mentioned before, we look at the different tags on Stack Overflow as subcommunities of the platform. Figure 12 shows the average post count until transition for the tags Go, VBA, Swift,

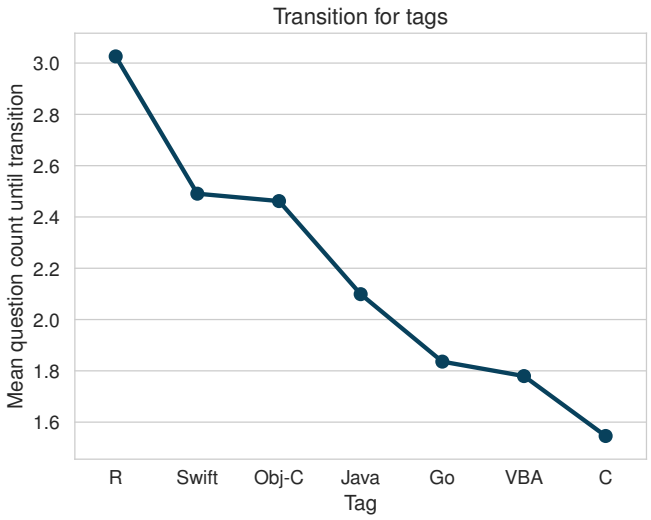


Figure 12: *Mean question count until tag-specific transition in decreasing order for the programming language tags Go, VBA, Swift, Objective-C, R, C and Java.*

Objective-C, R, C and Java. It is clear that the various tags also show vastly different averages

regarding transition. Therefore, we want to take a more proper look at the communities and what may cause these differences. We focus on the social relation between users and how tags relate to each other. For that, we are going to look at the social alignment of communities, their negativity and their social distance to one another.

2.3.1 Social Alignment

A Stack Overflow subcommunity does not necessarily reflect the whole subcommunity as not all of its members are active on the website. Most members may exclusively appear as visitors, looking up questions and answers, but not posting and voting themselves. When talking about the transition of users, we specifically only look at users registered on the platform. According to the co-founder and former CEO of the website, Joel Spolsky, Stack Overflow attracts around 100 million monthly visitors, a number far exceeding the number of registered users[8]. This means most users of Stack Overflow are just "visitors" who do not participate directly. A Stack Overflow subcommunity with a small share of those visitors is a tightly-knit group of people. This closeness may result in more positive interaction between the members, but could also appear as more closed-off to newcomers. Therefore, we consider the representativeness of a community in our model. For that we define a social alignment (SA) score. As we possess no information about the users that are only visiting

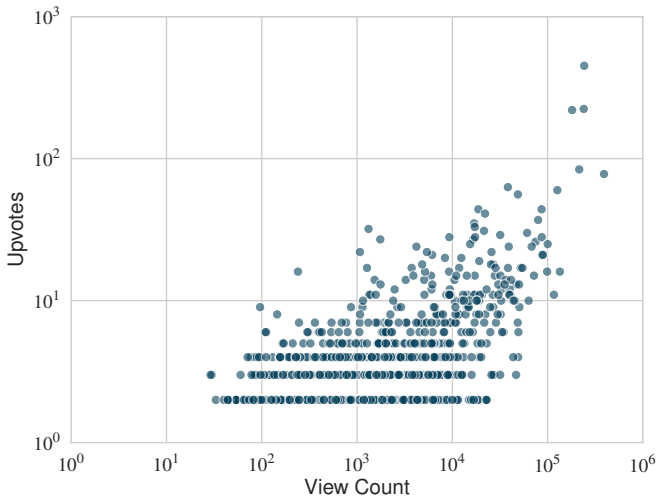


Figure 13: *Scatterplot of view count to upvote count for a random sample of 10,000 posts on Stack Overflow on a log-log scale. We remove points for posts with zero upvotes from the visualization. The plot highlights the positive correlation between those variables.*

the website in our dataset, another method of obtaining a representativeness score is needed. By looking at view counts of posts in relation to votes, which can only be cast by registered users who have already made a successful post, we gain some insight into how many of the community members participate on Stack Overflow. Furthermore, it shows to what degree the core members (the upvoters) of a Stack Overflow community influence questions gaining attention by the rest of

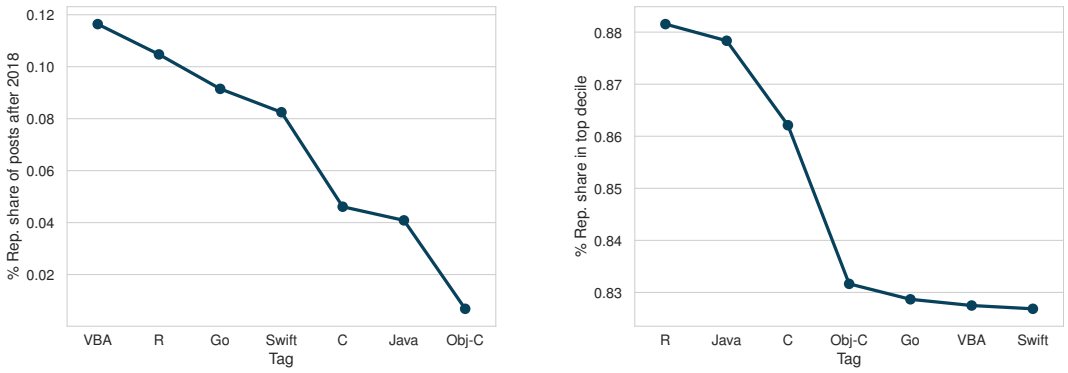
[8] https://www.joelonsoftware.com/category/reading-lists/stack-overflow/

the community. Therefore, we want to consider at how the view and upvote counts are correlated by calculating the Spearman correlation between post views and upvotes. Figure 13 depicts the relationship between the number of views and upvote counts for all question posts on a log-log scale.

With rg_X as the rank scores of values X, $Cov(rg_X, rg_Y)$ as the covariance of the rank variables and σ_{rg_X} as the standard deviation of X we calculate SA as:

$$SA = \rho = \frac{Cov(rg_X, rg_Y)}{\sigma_{rg_X}, \sigma_{rg_Y}} \tag{1}$$

With the SA score defined we can analyze what affects social alignment in communities and in which ones we do expect to see high or low scores. Communities change over time, for example in a linguistic way as new language norms arise. With such changes, users that have been in the involved for a long time can become alienated [13]. Generalizing this observation, we hypothesize that a tag with an over the years well-established core of users is more susceptible to developing division between its members. To test this hypothesis, we create an OLS regression model with SA as the dependent variable with five predictor variables. The first two of those are the question and answer count in a community. Additionally, we include the number of users who have created at least one post within that tag. Those measures should provide an estimation of the size of a community. Another variable to include is the percentage share of reputation gained in the community recently, more precisely since 2018. A substantial share indicates a tag is either new or has just become



(a) (b)

Figure 14: *Percentage share of reputation gained in a community since 2018 (a) / gained by the top reputation decile (b) for the tags VBA, R, Go, C, Swift, Java and Objective-C.*

popular in recent times. Such a tag may lack the aforementioned established core of users. The last variable included denotes the percentage share of reputation gained by the top reputation decile. A larger share indicates the community has a set of elite users collecting the most reputation. As the data set does not include information about the reputation gained from a single post, we calculate an estimation for that value as the difference between ten times the number of upvotes and two

times the number of downvotes of a post. This is an estimation of the method Stack Overflow uses to calculate reputation[9].

Figure 14 depicts the percentage share of reputation gained since 2018 and by the top reputation decile users for different tags. Looking at the evolution of the TIOBE index, a measurement of programming language popularity, the results regarding the reputation gained since 2018 seems to match the general trend in language popularity, with C, Java and Objective-C being on the decline and R, Go and Swift rising[10]. Regarding the share at the top decile, there appears to be significant differences between the tags. Table 3 shows the results of this regression model. We

Table 3: *OLS regression models predicting social alignment and community negativity.*

	Dependent variable:	
	Social alignment	Community negativity
	(1)	(2)
Intercept	0.434^{***}	0.057
	(0.065)	(0.044)
Question count (log)	0.043^{***}	-0.061^{***}
	(0.016)	(0.011)
Answer count (log)	-0.063^{***}	0.068^{***}
	(0.019)	(0.013)
User count (log)	0.001	0.011
	(0.011)	(0.007)
% of Rep. of Community in prev. year	-0.113^{**}	0.173^{***}
	(0.046)	(0.032)
% of Rep. of decile users in community	0.341^{***}	-0.246^{***}
	(0.086)	(0.059)
Observations	500	500
Adjusted R^2	0.060	0.136
F Statistic	7.165^{***}	16.75^{***}
Significance thresholds:	*p<0.1; **p<0.05; ***p<0.01	

see that the recent popularity of a community matters in its social alignment. Younger, rising ones appear to have lower social alignment than older ones without any significant rise in popularity in recent years, confirming our previous hypothesis. One possible explanation for this is that those younger tags and those that only recently have become popular are yet to develop a robust set of core contributors setting the direction of the community. The results also show that a community with a significant share of its reputation collected at the top decile have appear to have higher social alignment scores. The adjusted R^2 for the model is quite low, though.

2.3.2 Community Negativity

Many users have been upset with the hostility of some of Stack Overflow's users [37]. While one can look at the negativity of the website as a whole and analyze reasons for that on a macro-scale, we want to focus on the subcommunities of the platform and how users in those interact regarding positivity/friendliness. We expect to observe a difference in negativity between the subcommunities

[9] https://stackoverflow.com/help/whats-reputation
[10] https://www.tiobe.com/tiobe-index/

of Stack Overflow. For example, the programming language Ruby claims to “[have] a vibrant and growing community that is friendly towards people of all skill levels” [11], which may reflect on the Ruby tag on Stack Overflow. In contrast to that, Javascript with the various languages that compile to it like TypeScript and CoffeeScript and the many frameworks may show some division. Such a division could result in animosity regarding “splitters". To look at the negativity of each tag, we use a score called *community negativity* (CN). We define community negativity as the number of downvotes divided by the number of all votes cast inside a subcommunity:

$$CN = \frac{\#downvotes\ in\ subcommunity}{\#votes\ in\ subcommunity} \tag{2}$$

A high CN score indicates that negative feedback is more common in a community. We hypothesize that users engaging in negative tags are more afraid of answering questions as they may feel unwelcome. This would cause slower transition and especially lower transitioning probability.

As done with social alignment, we create a OLS regression model with CN as the predictor variable. Table 3 shows the results of that. We observe, that more recent tags show an increased negativity. In addition, communities with a significant reputation accumulation at the top seem to be less negative. This may stem from the fact that the top reputation users have a higher influence on the communities reputation while simultaneously being less downvoted in general.

2.3.3 CN/SA-Space

We want to take a more proper look at how social alignment and community negativity relate to each other and use both measures to discover clusters of communities. For that, we apply the k-means clustering algorithm from Python’s Sklearn library with $k = 4$ to our datapoints. Figure 15 shows the results of applying the algorithm. On average, the CN scores of the data points are decreasing and the SA scores increasing from cluster 1 to 4. We observe a clear correlation between social alignment and negativity, where less socially aligned communities are more negative. More precisely, the Pearson correlation coefficient between SA and CN is $\rho = -0.84$. The most visited tags of the first cluster, which on average has the lowest SA and the highest CN scores, consist of foundational technologies related to web development such as PHP or HTML. This cluster is the most spread out, consists of the least datapoints and has the highest average ranking of the tags ranked by the number of posts. It borders cluster 2, which consists of JavaScript and the related jQuery, but additionally includes programming languages often used in application development like C++ and Java. Cluster 3 shows some major programming language tags like C# and Python, but also shows mobile technologies with Android and ios. Cluster 4 has the lowest average CN and the highest SA score. It includes technologies related to web development that are not based on JavaScript (Ruby-on-rails, Django) and technologies related to Apple’s own infrastructure with iPhone and Xcode.

We utilize the SA and CN measures to analyze how communities change over time on Stack Overflow. The data for votes in our dataset contain the date at which a vote was cast. We can use

[11] https://www.ruby-lang.org/en/community/

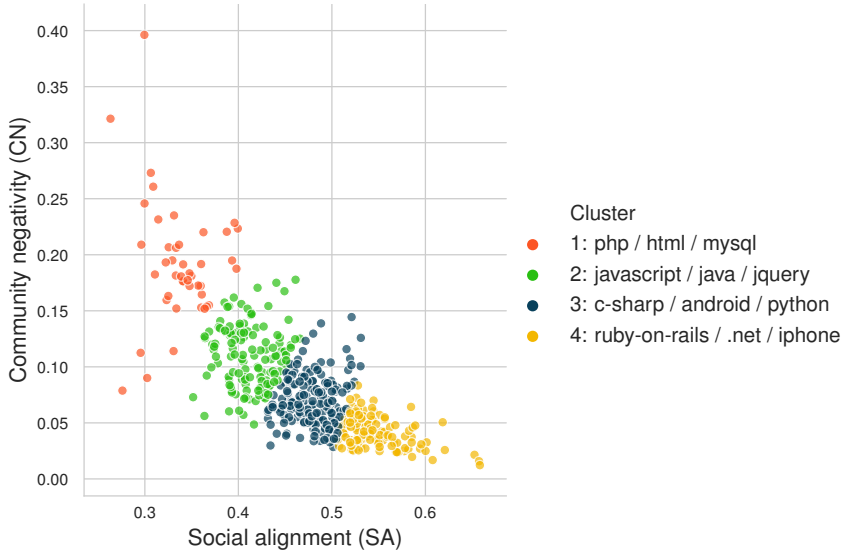


Figure 15: *Scatterplot of Social Alignment (SA) and Community Negativity (CN) scores for the top 500 communities (by number of posts) on Stack Overflow with k-means applied as a clustering method using k=4. On average, the CN score of the datapoints are decreasing and the SA scores increasing by cluster (from 1 to 4).*

those dates to calculate the CN score of a community at any time. The calculation of SA is a bit more difficult though, as it requires the view counts for each post for all the times we want to look at. This is unprovided in our data. Baltes et al. have created a dataset called *SOTorrent* combining the information of multiple Stack Overflow data dumps from various dates [5]. This makes it possible to analyze the evolutions of posts on the website over time. We utilize this information to get the view counts of posts at different snapshots. As the dataset does not contain pre-2016 information of Stack Overflow, we want to look at the evolution from 2016 to 2019. Figure 16 shows the progression of SA/CN for a selection of tags over this time period. All tags depicted show a decrease in social alignment over the years. There appears to be a growing disconnect between the voting users and the viewing public. There is not such a definitive change in community negativity values, though. The community negativity of VBA, Go and Swift has increased from 2016 to 2019, while it has decreased for Objective-C, R, C and Java.

2.3.4 Community Distance

Another interesting aspect to look at when considering different subcommunities is their distance to each other. We define two measures of community distance. The first one is the *community culture difference* (CCD), which we define as the Euclidean distance between two points $p = (SA_1, CN_1)$ and $q = (SA_2, CN_2)$ in the SA/CN space:

$$CCD = \sqrt{(SA_2 - SA_1)^2 + (CN_2 - CN_1)^2} \tag{3}$$

Table 4: *Table showing both the most top tags (most popular tags by the number of posts) and the center tags (tags closest to the centroid) of all four clusters shown in Figure 15 The tags are shown in order from most popular/closest from right to left. Cluster 0 appears to have many tags associated that relate to large-scale (professional) code projects. Cluster 1 contains basic web development tags, group 2 contains more advanced web technologies with a focus on Python/Ruby as programming languages, while cluster 3 also contains advanced web development technologies but with a focus on Javascript.*

Cluster 1					
Top Tags	php	html	mysql	arrays	regex
Center Tags	excel	python-3.x	web-scraping	unity3d	awk
Cluster 2					
Top Tags	javascript	java	jquery	c++	css
Center Tags	dictionary	outlook	replace	sql-server	batch-file
Cluster 3					
Top Tags	c#	android	python	ios	r
Center Tags	bitmap	twitter-bootstrap-3	plugins	azure	asp.net-mvc-4
Cluster 4					
Top Tags	ruby-on-rails	.net	iphone	django	xcode
Center Tags	exception-handling	testing	junit	https	.net

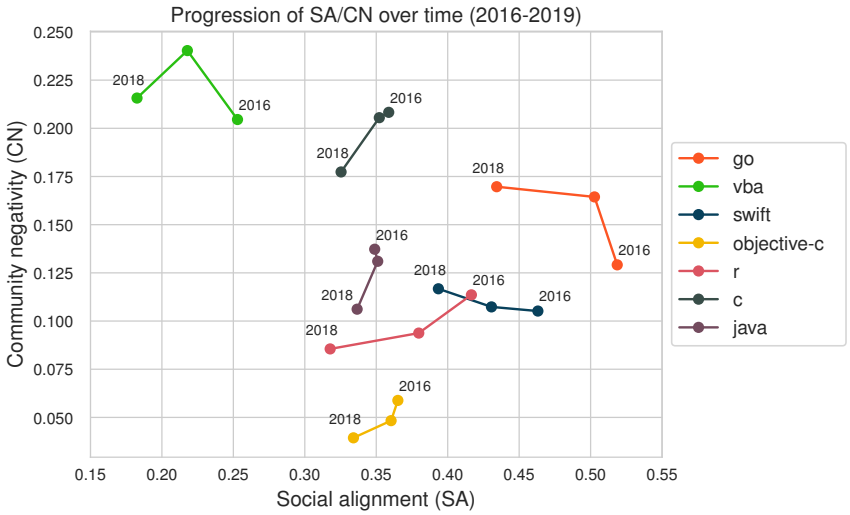


Figure 16: *Progression of social alignment (SA) and community negativity (CN) scores over the years 2016 to 2019 for a selection of Stack Overflow tags. The progression takes into account views and votes in the time periods in between 2016-09-12, 2017-06-12, 2018-09-05 and 2019-09-04 (ISO 8601)*

Another way of obtaining a distance measure between communities is to generate a network with tags as nodes, which are connected if they appear simultaneously in any post on Stack Overflow. We look at the top 500 most popular tags on the website according to the number of posts. Let $Tags$ be a set of those top 500 tags, which is the node set of our network. For $tag_i, tag_j \in Tags$, $i \neq j$ we define edges $(tag_i, tag_j) \in E$ if tag_i and tag_j appear together in one post. Each edge (tag_i, tag_j)

has a weight w_{ij} defined as:

$$w_{ij} = \frac{1}{\#posts\ where\ tag_i\ and\ tag_j\ appear\ together} \tag{4}$$

We filter the network generated in this way with the disparity filter algorithm [30]. We define the distance between two tags as the shortest path length between them. For two unconnected tags, the distance is set to the diameter of the network. We hypothesize that users with previous experience in a community with a smaller network distance to the new one transition faster and more likely in the new tag. The reasoning behind that is their increased familiarity with the website and similar technologies. More specifically, we expect experience with tags that are close to Swift to reduce the number of posts until transition. One example would be Objective-C which, as stated previously, has similarities and shares context with Swift [19].

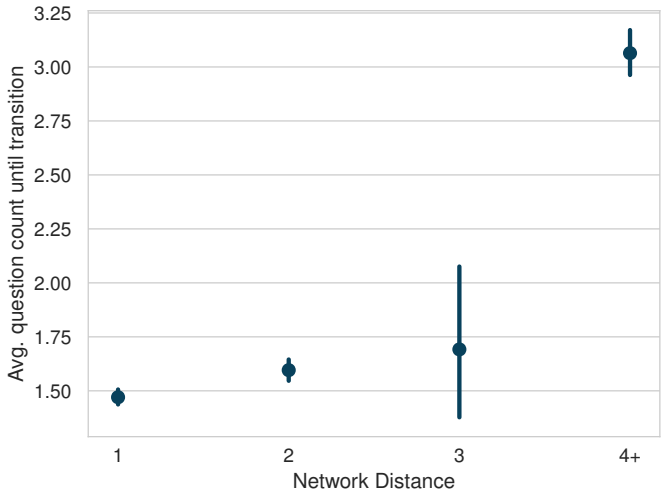


Figure 17: *Means of posts until Swift transition by the distance between Swift and their previous transitioned-in tagfor all users that do transition in the Swift subcommunity at some point in time. Users that do not have a previous transition are put into the 4+ category.*

2.4 Site-level Features

Figure 1 depicts how the share of answerers on Stack Overflow decreased over the years. We also know that the website has gone through multiple design changes which may have affected the user[12]. One of those changes happened in 2009-2010 when the amount of reputation gained from upvoted questions was reduced from ten to five. This was done to encourage users to provide answers instead of questions for gaining reputation [26]. Additionaly, as stated in the introduction of this thesis, the growth in popularity of Stack Overflow has changed what kind of individuals use the platform. Complaints about the drop in quality of questions have arisen, for example. Anecdotal evidence shows an increase in *noobs*, users that post trivial low-quality questions, clogging the

[12] https://meta.stackexchange.com/questions/59445/recent-feature-changes-to-stack-exchange

system. Additionally, there appears to be an increase in *help vampires*, users only asking questions, not trying to find the required knowledge on their own [34]. Therefore, it is likely that the years of posts and the age of user accounts are related to the posts until transition. We hypothesite that users with younger accounts take more posts until transition and are less likely to make the shift at all.

In the user data of our dataset, we are provided with the account creation dates all converted to UTC timezone. Stack Overflow was created in 2008, and the data contain posts created until the middle of 2019. Both the years 2008 and 2019 are not covered completely. For the year 2008, the reason for that lies in the fact that the platform was created in April of 2008[13], therefore there is no data for the pre-April months. Figure 18 depicts the mean post count until transitition by user

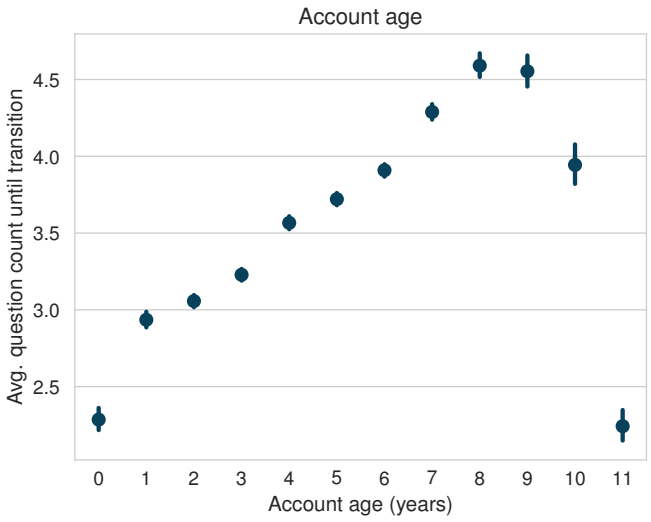


Figure 18: *Distributions of question count until transition, excluding outliers, of users by the year of account creation from 2008 to 2019. The means were increasing until 2011 to a value of around 4.5, from where they were continuously decreasing to a value of about 3 in 2018. For 2008, 2009 and 2019 the median is 1, in all other years it is 2.*

account age in years. We observe an increase in means in accounts age 11 to 8, peaking at around 4.6 questions until transition. From there on, the means are decreasing for each year. Between users with age 11 and 10, we observe a striking increase in means. We want to test the null hypothesis for all these previously mentioned binomial variables. To do that, we apply the Mann-Whitney U test. Table 5 displays the results of this. It can be seen that all those differences in transition regarding those categories are statistically significant.

3 Methods

With all these different variables that relate to the transition of a user defined, we want to generate a model for visualizing the distribution of the question count until transition. As we see from the

[13] https://blog.codinghorror.com/introducing-stackoverflow-com/

Table 5: *Results of the Mann-Whitney U-test of the number of questions until transition applied to binary user features categories showing both the values for statistics and p.*

Category	*Statistics*	*p*	*Avg. post count until transition*
AboutMe (Filled-out/Empty)	8.8×10^{10}	0	Filled-out: 3.40 Empty: 4.14
Website URL (Filled-out/Empty)	7.4×10^{10}	0	Filled-out: 3.13 Empty: 4.14
Gender (Male/Female)	2.3×10^{9}	<0.05	Male: 3.51 Female: 4.16
Geography (Global North/Global South)	2.1×10^{10}	<0.05	North: 3.39 South: 3.43
Weekender (Yes/No)	5.1×10^{10}	<0.05	Weekender: 3.74 Weekdayer: 3.88

distribution in Figure 3, standard linear regression, which assumes that the dependent variable is continuous and unbounded, is not an appropriate model choice. Therefore, we need to look at a different regression model. For that we are going to focus on negative binomial and logisistic regression.

3.1 Negative Binomial Regression

We can define the posts of a user as a Bernoulli trial with the two outcomes question and answer. When looking at questions as "successes" and answers as "failures", the probability distribution for the question count until transition is the negative binomial (NB) distribution. This is the distribution of the number of successes before a specific number of failures occur in a sequence of independent Bernoulli trials [22]. With r as the number of failures, which each have a probability of p, and k as the number of successes with probability $(1-p)$, the probability to see a sequence of length $n = r+k$ with exactly r failures and k successes is $(1-p)^r p^k$. We know that the r-th failure has to occur at the end of the sequence. Therefore, the k successes have to happen in the remaining $k+r-1$ trials. The number of sequences fulfilling these criteria is $\binom{k+r-1}{k}$. Thus, with $r,k \in \mathbb{N}_0$, the probability mass function of the NB distributions is defined as:

$$P(X = k) = \binom{k+r-1}{k}(1-p)^r p^k \tag{5}$$

As the binomial coefficient does not take real numbers as input, we cannot use this definition of the distribution with $r \in \mathbb{R}$. To make this possible, one can utilize the Gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ which has the probabiity $\Gamma(x) = (x-1)!$. Inserting Γ into equation (5) results in:

$$P(X = k) = \frac{\Gamma(r+k)}{k!\,\Gamma(r)}(1-p)^r p^k \tag{6}$$

Furthermore, Hilbe utilizes the mean of the distribution $\mu = \frac{(1-p)r}{p}$ for the specification of the distribution [22]. This expression of μ leads to an expression of p as $p = \frac{r}{\mu+r}$ and $(1-p)$ as

$1-p=\frac{\mu}{\mu+r}$, resulting in a specification of the NB distribution as:

$$P(X=k)=\frac{\Gamma(r+k)}{k!\Gamma(r)}\left(\frac{r}{\mu+r}\right)^{r}\left(\frac{\mu}{\mu+r}\right)^{k} \tag{7}$$

for $r\in\mathbb{R}, k\in\mathbb{N}_0$. The corresponding regression model for this distribution is:

$$\ln(\mu)=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p \tag{8}$$

with the regression variables $x_1,...,x_p$ and the coefficients $\beta_0,...,\beta_p$ that are to be estimated. To fit a negative binomial model, maximum-likelihood estimation is applied with the goal of learning a estimate parameter vector β and $\alpha=\frac{1}{r}$. For a sample of n datapoints with the dependent variable of the point i as y_i and the predictor variables as x_i, we can write the maximum likelihood equation using exponention on equation 7 as:

$$l(\alpha,\beta)=\prod_{i=1}^{n}\frac{\Gamma(y_i+1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha e^{x_i\cdot\beta}}\right)^{1/\alpha}\left(\frac{\alpha e^{x_i\cdot\beta}}{1+\alpha e^{x_i\cdot\beta}}\right)^{y_i} \tag{9}$$

The corresponding log-likelihood function is:

$$\begin{aligned}L(\alpha,\beta)=\sum_{i=1}^{n}(y_i\ln\alpha+y_i(x_i\cdot\beta)-\left(y_i+\frac{1}{\alpha}\right)\ln(1+\alpha e^{x_i\cdot\beta})\\+\ln\Gamma\left(y_i+\frac{1}{\alpha}\right)-\ln\Gamma(y_i+1)-\ln\Gamma\left(\frac{1}{\alpha}\right))\end{aligned} \tag{10}$$

To maximize α and β, one has to apply a iterative algorithm, for example the expectation-maximization algorithm proposed by Adamidis et al. [1]. As the negative binomial model is non-linear, one cannot calculate a R^2 as in linear regression model. To still obtain some information about goodness of fit for such a regression model, many ways of calculating a "pseudo" R^2 value have been defined. One of those is the McFadden Pseudo-R^2:

$$R^2_{McFadden}=1-\frac{L}{L_0}=1-\frac{Deviance}{NullDeviance} \tag{11}$$

with L_0 as the log-likelihood and *NullDeviance* as the deviance of the null model.

3.2 Logistic Regression

As mentioned before, we have to deal with selection bias in our data for predicting the post count until transition as we have many users not transitioning, which we need to drop from the dataset for the NB regression. To account for this bias, we include a model to predict whether a user transitions or not. As this is a dichotomous variable, we can apply logistic (logit) regression for this task [23].

Figure 19 shows a comparison between linear and logistic regression applied to a randomly genrated dataset with a binary dependent variable. By looking at the two regression results, we can already see that the logistic regression one is more reasonable then the linear one. One reason for

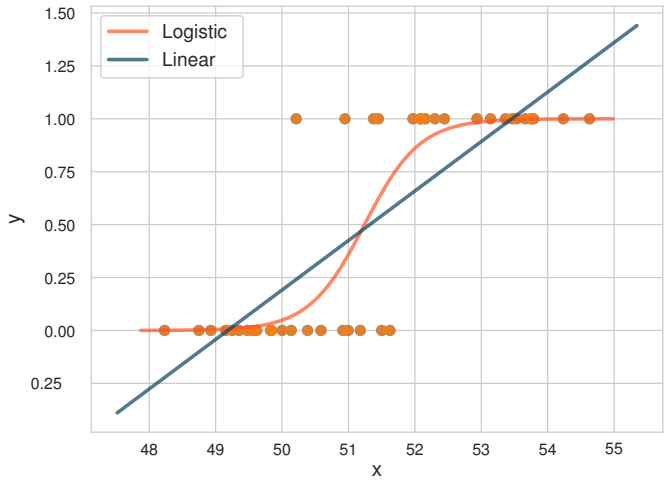


Figure 19: *Comparison of the results from logistic and linear regression on randomly generated data with a dichotomous dependent variable y. The logistic regression returns predictions between 0 and 1, representing a likelihood.*

this is that using linear regression does not limit the range of the dependent variable. As we want to predict probabilities, a prediction of $x \notin [0,1]$ does not make any sense. Therefore, the regression results need to be mapped to $[0,1]$. In logistic regression, this is achieved by applying the *logistic function* (also called sigmoid function) $\sigma : \mathbb{R} \to [0,1]$:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{12}$$

Considering the typical linear equation

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{13}$$

and applying $\sigma(x)$, we get a variation of the equation:

$$p(x) = \sigma(\hat{y}) = \frac{1}{1+\exp\left(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)\right)} \tag{14}$$

Applying the inverse of σ, we get the final equation used in logistic regression:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{15}$$

The left-hand side of equation (15) is the log-odds of the classification x. This means that logistic regression predicts the log odds of a classification with $odds = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}$. To fit a logistic regression mode, maximum-likelihood estimation is applied [2]. The goal here is to learn a estimate parameter vector $\hat{\beta}$. Considering n samples, labeled either 0 or 1, we want $\hat{\beta}$ to be learned in such a way that $\widehat{p(x)}$ (for samples labeled 1) and $1-\widehat{\beta}$ (for samples labeled 0) to be as close to one as possible. With y as the label vector, these two conditions result in the

likelihood-function:

$$\begin{aligned} l(\beta) &= \prod_{i,y_i=1} p(x_i) \times \prod_{i,y_i=0} (1-p(x_i)) \\ &= \prod_i p(x_i)^{y_i} \times (1-(p(x_i))^{1-y_i} \end{aligned} \tag{16}$$

This likelihood can be transformed into the *log-likelihood* [23], resulting in the equation:

$$\begin{aligned} L(\beta) &= \sum_i^n y_i \ln(p(x_i)) + (1-y_i)\ln(1-p(x_i))) \\ &= \sum_i^n y_i \ln\left(\frac{1}{1+e^{-\beta x_i}}\right) + (1-y_i)\ln\left(\frac{e^{-\beta x_i}}{1+e^{-\beta x_i}}\right) \\ &= \sum_i^n y_i \left[\ln\left(\frac{1}{1+e^{-\beta x_i}}\right) - \ln\left(\frac{e^{-\beta x_i}}{1+e^{-\beta x_i}}\right)\right] + \ln\left(\frac{e^{-\beta x_i}}{1+e^{-\beta x_i}}\right) \\ &= \sum_i^n y_i \left[\ln\left(e^{\beta x_i}\right)\right] + \ln\left(\frac{1}{1+e^{\beta x_i}}\right) \\ &= \sum_i^n y_i \beta x_i - \ln(1+e^{\beta x_i}) \end{aligned} \tag{17}$$

The goal of the maximum-likelihood estimation is to find β such as to maximize $L(\beta)$:

$$\beta = \underset{\beta}{\operatorname{argmax}} L(\beta) \tag{18}$$

As $L(\beta)$ is a transcendental equation, one has to use numerical methods to get an estimation of the coefficients. Popular methods for this are the Newton-Conjugate-Gradient and the limited-memory BFGS, which is the default estimator used for logistic regression in the scikit-learn library[14].

4 Results

We utilize a negative binomial regression model to predict the number of questions until transition for users who do make the shift. As mentioned previously, there are some individuals that do not transition or have not done so yet. To account for these users, we fit a Logit regression model to predict whether a user transitions at all. Besides the general case, we look at the subcommunity of the programming language Swift. For all models used, we utilize two different datasets for fitting. Besides using data containing all the features mentioned previously, we additionally use a dataset unfiltered in regards to gender and geography features. The latter dataset contains significantly more observations.

[14] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

4.1 General

To generate a negative binomial regression model, we use the GLM (Generalized Linear Models) submodule from the statsmodel package in Python[15]. For all models, an intercept term is included. As predictive variables for the regression, we implement the features we introduced previously. Regarding the profile page, we include two boolean variables; one to encode a user having a filled-out profile page and one to encode the profile containing a link to a website. Another variable encodes the person's gender classification (0: Male, 1: Female). For the geography feature, we include a variable set to one if the user appears to be from the global north. Another dichotomous variable is included to encode a individual being a weekender or not.

Regarding tag-level features, we aim to include our measure of community negativity into the regression. To be able to include tag-level features into the model, we look at the subcommunities each person posts in. We define specialization by looking at the most frequently used tags by a user pre-transition (including the transition post in the calculation). We say a tag is the person's *specialty* if it is the most frequently appearing one in pre-transition posts. A user may have multiple specialties if there is a tie in most frequent tags. We assume the specialties of a user to have the most significant effect of all their subcommunities on transition. Specialties are used to assign values for the tag-level features. If a user only has one specialty, that tag is used to get the appropriate feature value. When looking at an individual with only Python as a specialty, for example, the value for community negativity of that person is set to the CN score of the tag Python. On the other hand, if an individual has multiple specialities, the average of all values is picked. If the user from the previous example also has a specialty in Java, the community negativity matching to that user is set as the average of CN scores of Python and Java. When looking at specialty, only the top 500 tags by question count are considered for size reasons. If an individual does not have a specialty that is also a top 500 tag, we say that their specialty is "other". The related community variables for users with such a specialty are set to the site-wide value for SA, CN and the reputation shares and to the overall community average for the tag's user count. With this definition, each user can be assigned tag-level features highlighting the influence of the tags they previously interacted with. Additionally, we add the specialties of each user as dummy variables to the NB regression model.

As social alignment and community negativity are strongly correlated, we choose to exclude social alignment from the regression. Additionally, we include variables for the community size, the total share of the communities reputation gained in 2018 and the total share of community reputation gained by the users in the top decile of reputation gained. A considerable share of the total reputation gained in 2018 shows a community to either be new or to only be popular recently. The reputation share in the top decile serves a similar purpose as social alignment in that it shows whether there is a group of core users in the community. Such a group is likely to gain a significant share of reputation points. We use the same variables for the Logit regression model with one change to the definition of specialty: if a user does not transition, all posts made by that user up until the most recent point in the data are considered in the specification of the specialty. Considering site-level features, a variable for account age is included.

[15] https://www.statsmodels.org/stable/index.html

Table 6 shows the results of the NB/Logit regression models for regarding general questioning-to-answering transition. Both models are trained on data excluding gender and geography information (1) and data including those (2).

Table 6: *General transition regression results. For both models, the first column is the model, including all users without filtering for gender and geography. The second column includes both inferred gender (woman = 1) and geography (north/south).*

	General transition			
	Questions until transition (NB)		Has transition (Logit)	
	(1)	(2)	(1)	(2)
Intercept	-8.029^{***}	-7.503^{***}	2.493^{***}	1.629
	(0.064)	(0.148)	(0.126)	(0.403)
Account age	0.048^{***}	0.050^{***}	0.165^{***}	0.258^{***}
	(0.003)	(0.001)	(0.001)	(0.003)
Website URL on profile	-0.200^{***}	-0.133^{***}	0.957^{***}	0.542^{***}
	(0.003)	(0.007)	(0.008)	(0.018)
Filled-out AboutMe	-0.080^{***}	-0.062^{***}	1.305^{***}	0.953^{***}
	(0.003)	(0.006)	(0.006)	(0.015)
Inferred Woman		0.115^{***}		-0.727^{***}
		(0.012)		(0.023)
Frequent Weekend Poster	-0.025^{***}	0.004	-0.198^{***}	-0.174^{***}
	(0.003)	(0.008)	(0.005)	(0.017)
From global north		-0.047^{***}		-0.091^{***}
		(0.007)		(0.016)
Community negativity	-0.935^{***}	-0.717^{***}	-0.370^{**}	-0.146
	(0.090)	(0.208)	(0.171)	(0.543)
Users in community (log)	0.222^{***}	0.227^{***}	-0.269^{***}	-0.148^{***}
	(0.001)	(0.010)	(0.002)	(0.007)
% of Rep. of Community in Prev. Year	1.019^{***}	0.979^{***}	0.012	0.038
	(0.092)	(0.208)	(0.157)	(0.442)
% Rep. of top decile users in community	7.335^{***}	6.612^{***}	0.791^{***}	0.797^{*}
	(0.072)	(0.166)	(0.139)	(0.441)
Observations	813,365	155,263	1,188,415	183,812
Pseudo R^2	0.353	0.363	0.205	0.213
Significance thresholds:	*p<0.1; **p<0.05; ***p<0.01			

4.2 Swift

Besides the features included in the general models, we include variables for previous experience and the distances between Swift and the pre-Swift tags of users. To include such information, we introduce a variable for the question count of users before they enter the Swift subcommunity and a dichotomous variable set to one if the user transitions in another subcommunity before joining Swift. The model also includes both a variable for the community culture difference and network

distance. From all tags an individual previously posted in, we only select the one with the shortest distance to Swift as the second tag in calculating the distance measures. All network distances larger than or equal to four are binned together. Table 7 shows the results of the NB/Logit regression models questioning-to-answering transition for the tag Swift. Once more, two datasets are used for fitting, similar to the regression models for general transition.

Table 7: *Swift transition regression results. For both models, the first column is the model, including all users without filtering for gender and geography. The second column includes both gender (female=1) and geography (north/south).*

	Swift transition			
	Questions until transition (NB)		Has transition (Logit)	
	(1)	(2)	(1)	(2)
Intercept	0.886^{***}	0.525	-0.876^{***}	-0.082
	(0.084)	(0.174)	(0.121)	(0.269)
#Questions before first Swift post (log)	0.182^{***}	0.174^{***}	-0.572^{***}	-0.557^{***}
	(0.005)	(0.011)	(0.008)	(0.018)
Has previous transition	-0.479^{***}	-0.429^{***}	0.191^{***}	-0.099
	(0.042)	(0.091)	(0.060)	(0.139)
Account age	-0.031^{***}	-0.037^{***}	0.099^{***}	0.120^{***}
	(0.004)	(0.008)	(0.005)	(0.012)
Website URL on profile	-0.037^{**}	0.003	0.202^{***}	0.100^{**}
	(0.016)	(0.031)	(0.025)	(0.049)
Filled-out AboutMe	-0.037^{**}	-0.013	0.764^{***}	0.601^{***}
	(0.014)	(0.031)	(0.021)	(0.046)
Inferred Woman		-0.072		-0.046
		(0.058)		(0.085)
Frequent Weekend Poster	0.079^{***}	0.008	-0.232^{***}	-0.182^{***}
	(0.017)	(0.036)	(0.022)	(0.053)
Distance (Swift-prev. Tag)	0.147^{***}	0.151^{***}	-0.553^{***}	-0.525^{***}
	(0.015)	(0.033)	(0.023)	(0.052)
From global north		0.096^{***}		-0.239^{***}
		(0.033)		(0.050)
Community culture difference	-0.447^{***}	-0.045	4.407^{***}	3.756^{***}
	(0.124)	(0.249)	(0.192)	(0.410)
Observations	34,585	8,062	60,911	11,801
Pseudo R^2	0.243	0.249	0.187	0.138
Significance thresholds:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$			

5 Interpretation

Table 6 reports the coefficients of the regression models for general transition. We now interpret these coefficients. With each additional year in the user's account age, the number of posts until transition increases by $e^{0.048} - 1 \approx 0.049 = 4.9\%$ in the broader model excluding gender and geography data and 5.1% for the model including those variables. In the following discussion we report the effect sizes of the models on the narrower data set in parentheses. While it appears that users with older accounts take more questions until transition then ones with younger accounts, we also see that each additional year in age also increases the odds of making a transition at all by around 17.9% (29.4% for the model with gender and geography). This shows selection bias: newer users transition after fewer questions, but make the shift less often than older ones.

Users linking to an external website on their profile transition after $1 - e^{-0.200} \approx 0.181 = 18.1\%$ (12.5%) fewer posts. Additionally, for users with website links, the odds of transitioning increase by 160.4% (71.9%) compared to those without. Similarly, users with a filled-out About Me page transition 7.7% (6.0%) faster, while also having an increase in the odds of transitioning by 268.8% (159.3%). Having put more information on the profile seems to have a significant relationship with the chances a user has transitioned or not. Weekend users appear to transition after around 2.5% fewer posts than weekdayers when excluding gender and location. Being a weekender also decreases the odds of transitioning by around 18.0% (16.0%). This means that we observe a selection bias for the broader model, where weekenders do transition less often, but if they do, do so faster. An increase in the CN value by 0.1 is related to an decrease of questions pre-transition by around 8.9% (6.9%), and lower odds of transitioning by 3.6%. The decreased question count for more negative communities is an unexpected result. Users classified as likely women by our method transition around 12.2% slower than those classified as male, while also having 51.7% lower odds of transitioning at all. Users from the global north transition around 4.6% faster than those from the south, but their transitioning odds are 8.7% lower. Each increase in the log of community size by one increases the post count until transition by 24.9% (25.5%) while also showing decreasing transition odds by 23.6% (13.8%). Each increase in the reputation share in 2018 by 10 percentage points shows an increase in pre-transition question count by 10.7% (10.3%). Similarly, a one percentage point increase in the reputation share of the top decile increases post count until transition by 9.8% (6.8%). All mentioned relationships are statistically significant ($p<0.05$).

Table 7 shows the results of the Swift community analysis. Users filling out the About Me section are 114.7% (82.4%) more likely to transition. Users linking a website on their profile have a 3.6% faster transition compared to users with no website link. Their odds of transitioning are also 22.4% (10.5%) higher. Weekenders appear to transition less likely than weekday users. The odds of transitioning are 20.7% (16.6%) lower for the former group of users. Each one point increase in the log of the number of questions a user asks before their first Swift post increases the question count pre-transition by 20.0% (19.0%) while also decreasing transitioning odds by 43.6% (42.7%). Those odds are also decreased by 21.0% for users that have a subcommunity transition before

joining Swift. Users with previous transition show also show to transition after 38.1% (34.9%) fewer posts. Users from the global north appear to transition 10.1% slower and 21.3% less likely in the Swift tag, contrasting the results from the general transition model. For the tag-level measures we observe slower transitions for users with pre-Swift specialities that are further away from Swift in the tag network. Each one point increase in that distance increases the post count until transition by 15.8% (16.3%) while also decreasing transitioning odds by 42.5% (40.8%). Additionaly, the community culture difference shows to have an effect on transitioning odds. Those increase by 55.4% (45.6%) for a 0.1 point increase in CCD. An individual with such an increase in CCD also appears to transition after 36.0% fewer posts.

Each additional year in the user's age decreases the question count until transition by 3.1% (3.6%) while increasing transitioning odds by 10.4% (12.7%). All other results do not appear to be statistically significant, notably those regarding gender. These results show that previous experience on the website before stepping into the Swift community of Stack Overflow and how closely that previous experience relates to the new tag are the most significant factors in determining the question count pre-transition, while social factors do not seem to matter as much. Social factors, especially the profile information, seem to influence the transition odds, though.

6 Discussion

Our empirical findings suggest that there are several notable individual and community level factors which predict that some users take longer to become answer posters on Stack Overflow. The existence of significant differences between groups of users suggests that Stack Overflow's issues with representativity are more significant among its most active contributors.

Though our models do not describe cause and effect relationships, they do suggest areas for potential experimentation. While Stack Overflow cannot change, for example, the previous experience of a user, it may try to foster a more welcoming environment. We focus on the results regarding gender, geography and community negativity to propose possible changes to the website.

Mentor programs and other similar techniques for peer-guided assistance on the website have been shown to support new users integrate into the community faster [18]. Especially women appear to reengage sooner in the presence of such peers [16]. Those mentor programs usually only help new users formulate question posts. We propose that looking into such a program with an increased focus on the answering process may be a viable way to not only make questioning-to-answering transition less frightening, but also increase the number of answers in general. Such a system could have a mentor looking over a proposed answer by another person who wishes such assistance with the mentor gaining reputation points for their help. Opening up questions in that manner to editing (with the explicit wish of the posts creator) could also be made open to contribution by other less experienced individuals. This would allow new users to crowdsource an answer, splitting up the work to a degree that barriers to entry are decreased. Such an answer could present all the contributing users as post creators, dividing responsibility in case of negative feedback. Technically, such a way of contributing is already provided by Stack Overflow with the edit mechanism. In practice, only 12.6% of edits are from users that are not the initial post

creator, though [5]. Furthermore, surveys show especially women are sometimes not aware of Stack Overflow features like edits and badges [17]. Having mentors and crowdsourcing of answers be more present therefore is a key facet of the success of such programs. A simpler, hands-off idea would be to enable users to recommend other people which might be able to answer a question. This is already informally achieved by "paging" users in comments below questions.

The results show community negativity having an effect on transition. More negative communities appear to feature users with do not complete the questioning-to-answering transition. Over the years, users have started to complain about duplicate and trivial questions, for example from the previously mentioned help vampires [34]. Such posts are seen as annoying by the core of Stack Overflow users, which could cause negative reaction, downvotes and deletion. To reduce the negativity caused by such questions, a way of marking those posts as "beginner-friendly" could be introduced, similar to the "good first issue" label on GitHub[16]. Instead of downvoting or flagging a post for deletion, the more experienced users could vote on the post being moved into that category. They could then be highlighted to new users as possible entry points for contribution on the platform. Besides reducing negativity, such a label could therefore also reduce entry barriers for participation regarding previous programming experience.

To overcome language barriers, Stack Overflow has introduced variations of its platform in Japanese, Portuguese, Spanish and Russian. Those counterparts provide an opportunity for non-English users and those that are not so comfortable in the English language to still make use of the platform. One problem with the alternative platforms is the reduced participation and information present compared to the original Stack Overflow. A survey conducted by Botto-Tobar et al. shows how Portuguese speaking users still prefer the English website for this exact reason, while nevertheless acknowledging the beneficial effect of a reduced entry barrier [7]. Therefore, focussing on these non-English counterparts can provide a short-term solution for the problem, but other measures should be thought of in the future. Splitting up the Stack Overflow community into multiple websites and increasing isolation cannot be a desirable effect. Stack Overflow has been strict about its language policy [36], but it might be worth to reconsider that stance to some degree. A user could be provided with the opportunity to select languages on profile creation and to be shown posts in those languages. This would prevent questions in languages a user might not understand from flooding their questions page while still keeping non-native English users on the same platform. The amount of participation in non-normative language communities will still be lower than that of the English speaking part of the website, but measures like increased reputation gained from answering non-English posts and similar gamification strategies could foster participation. Eventually, machine translation may be leveraged directly on the platform.

7 Conclusion

We set out to design a model for the questioning-to-answering transition on Stack Overflow. We looked at different variables that may affect the question count until transition and compiled a list

[16] https://help.github.com/en/github/building-a-strong-community/encouraging-helpful-contributions-to-your-project-with-labels

of various hypotheses of how these effects might be expressed. With multiple models generated, we can evaluate the evidence for each hypothesis stated beforehand. Table 8 shows a summary of the main results of this thesis. At the individual level, we observe that the user's motivation behind using Stack Overflow has a significant effect on questioning-to-answering transition with individuals placing more effort into their profile, possibly for self-promotion, transitioning after fewer posts. Besides that, groups traditionally underrepresented on the platform like women and people from the global south face transition barriers. Additionally, community attributes like negativity and eliteness seem to hinder user transition. Previous activity in other subcommunities and the way

Table 8: *Summary of hypotheses and actual results for all considered features from the general transition regression models. Except for the values for gender and geography, the results are based on the broader NB/Logit model.*

Feature	Hypothesis	Result	Hypothesis confirmation
Profile page			
About Me	Users with filled-out About Me sections transition after fewer posts and more likely	7.7% fewer posts, 268.8% increased odds	Yes***/Yes***
Website URL	Users with website links transition after fewer posts and more likely	18.1% fewer posts, 160.4% increased odds	Yes***/Yes***
Gender	Woman take more posts until transition and transition less likely	14.0% more posts, 58.1% decreased odds	Yes***/Yes***
Geography	Users from the global north transition after fewer posts and more likely	4.9% fewer posts, 20.2% increased odds	Yes***/Yes***
Weekend posters	Weekenders transition after fewer posts but less likely	2.5% fewer posts, 18.0% less likely	Yes***/Yes***
Account age	Newer users take more posts until transition and less likely	4.9% more posts, 17.9% increased odds (with each one year increase)	Yes***/No***
Community features			
Community negativity	Users from negative communities take more posts until transition and do so less often	8.9% fewer posts, 3.6% lowered odds (with each 0.1 increase in CN)	No***/Yes**
User count in community	Users posting in larger communities take more posts until transition and so so less often	24.9% more posts, 23.6% decreased odds (with each increase of the log-size by 1)	Yes***/Yes***
Reputation share by top users	Users posting in communities with a concentration of reputation at the top take longer until transition and do so less often	9.8% more posts, 0.8% increased odds (with each 0.01 increase in rep. share)	Yes***/No***

the new and old communities relate to each other also plays a role in the questioning-to-answering shift. Regarding question count, the hypothesis that users with more questions asked pre-Swift would show faster transition has not been proven. One possible explanation is that individuals with a significant amount of questions asked pre-Swift are probably going to continue their behavior in the new subcommunity. We observe that the number of posts until transition and the probability for making the shift have changed over the years, which may relate to the growth and changes of the website.

Table 9: *Summary of hypotheses and actual results for all considered features from the Swift transition regression models. Except for the values for gender and geography, the results are based on the broader NB/Logit model.*

Feature	Hypothesis	Result	Hypothesis confirmation
Previous experience			
Question count	Users with more previous questions transition after fewer posts in a new community and do so more likely	20.0% more posts, 43.6% decreased odds (for one point increase in log)	No***/No***
Previous transition	Users with a previous transition take fewer posts until transition and are more likely to do so in a new subcommunity	38.1% fewer posts, 21.0% increased odds	Yes***/Yes***
Distance			
Network Distance	Users with prev. experience in communities further away from Swift transition after more posts and do so less likely	15.8% more posts, 42.5% reduced odds (for each distance point)	Yes***/Yes***
Cultural difference	Users with prev. experience in communities with higher cultural difference take more posts until transition and do so less likely	36.% less posts, 55.4% increased odds (for 0.1 increase)	No***/No***

7.1 Limitations

We have obtained results and used those to analyse our hypotheses, but there are some problems with the results. First off, as already mentioned, the dataset does not provide exact information about gender and location of the users, forcing us to infer those variables differently. The algorithms used for the classification of gender and location for this thesis are very basic, possibly causing many false negative classification. The model's accuracy is also decreased by selection bias. Using a Logit model predicting whether a user transitions or not, we have shown that selection bias does affect our NB model. Nonetheless, our NB model does not utilize a method for selection bias

handling. When talking about selection bias, one should also not forget that we only have access to data from registered users. Therefore, we defined the beginning of transitioning into adding to the collectivised knowledge of the platform with a user's first post. One could argue that it would be even more meaningful to define the start of this process as the first time a user interacts with Stack Overflow in any way. The limitation of the dataset used does not allow such a definition.

What we also have not looked into in this thesis is the possibility of editing posts. One flaw possibly caused by this is that a post's contents and tags can be changed after its creation. Changing tags can be done after amending a post to fit additional context. For example, there are Swift posts from 2008 in our dataset, even though Swift was released in 2016.

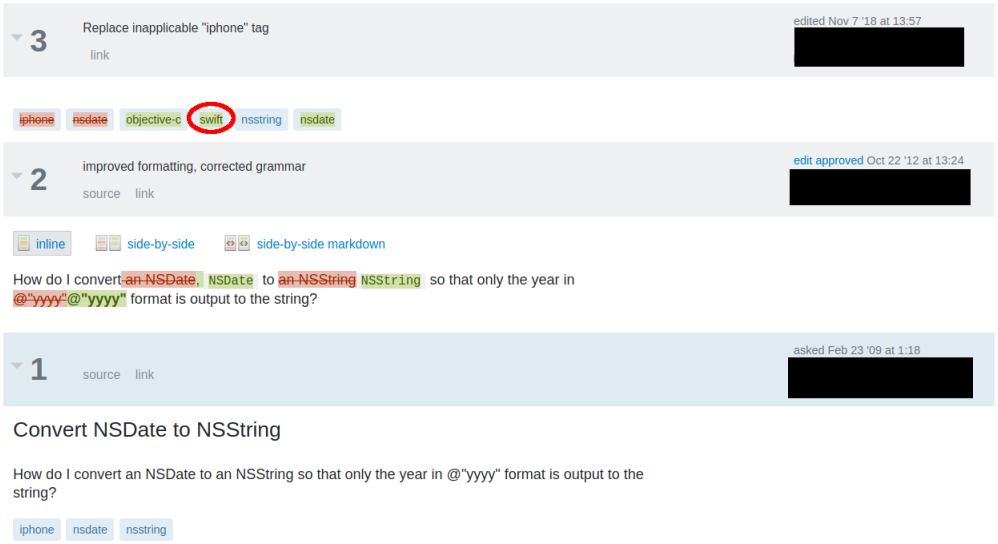


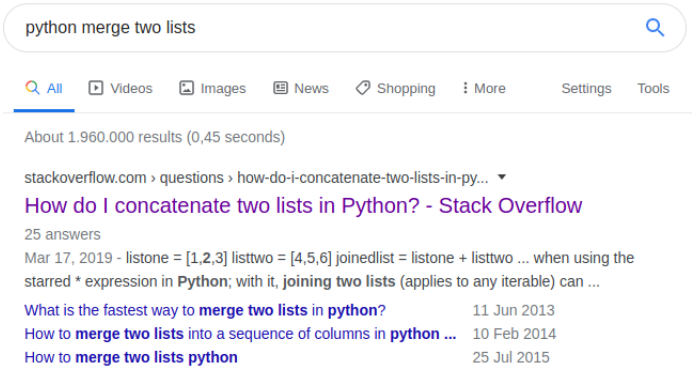
Figure 20: *Example of a post with edited tags. Here, the tag Swift is added in 2018 to a post initialy created in 2009. In the data dump, this post is shown to have the tag Swift and 2009 as the creation date.*

Those have been edited after 2016 to additionally provide context for the tag Swift. A tag can also be added because it was overlooked by the post's creator. Considering such an edit when looking at transition is the accurate way of doing it, but edits may still cause some impreciseness in the tag-related features.

7.2 Future work

In the future, one could include more variables into the regression models. For example, including a variable for the time of the day during which the user's posts may affect the regression. This variable would likely help to differentiate in a similar way to to the definition of a weekend poster. Someone who posts during daytimes is probably more likely to do so in a job-related context. Moreover, there is some room for improvement regarding the way gender and location are classified as we have used only basic methods to do so. For example, our gender classification cannot categorize users with nicknames as usernames. Only classifying users that have their first name in their username creates a selection bias: individuals who share this information in their username have to deal (and are able

to deal) with the social pressure accompanied by the disclosure of that information. Applying a more sophisticated way of classifying users, possibly utilizing profile pictures and time of posting, may provide more accurate results. The same applies to the location classification with cases not identified as correct locations by our algorithm.



(a)

(b)

Figure 21: *(a) Search result of the query "python merge two lists" on Google's search engine. (b) Search result of the query "python merge two lists" on the Stack Overflow search engine. Both search engines show different results for the same query.*

It would also be interesting to think about combining the NB and Logit model into one model directly accounting for selection bias. One way of doing so would be to apply a variant of the Heckman correction [21] utilizing Probit and NB regression. Roughly speaking, the Heckman correction is a two-stage model that first estimates the effects of selection bias (in our case on whether or not a user posts an answer at all) and then uses this estimate to adjust the second model (predicting the number of posts to transition). As we have merely looked at Stack Overflow posts that are still available, one could also think about including deleted posts into the calculations.

Getting many of their posts removed could be a factor for a slow transition or no transition at all of a user. To look at the negativity of a community differently than with our CN score, one could apply sentiment classification methods to infer positivity or negativity of a community. When analyzing the contents of posts, one may also look at the politeness model proposed by Danescu-Niculescu-Mizil et al. [12]. The politeness level of a community might affect transition.

It is highly likely that social feedback plays an important role in transition times. For example, a poster who receives a down-vote on his or her first question, may be less likely to quickly transition to posting answers. As social feedback is clearly confounded with the quality and past experience of the poster, we could not reasonably estimate the effect of feedback on time to transition. An idealized random experiment in which one randomly up or down votes a new poster's first question would be unethical. Nevertheless, this is clearly a topic of interest and deserves further attention.

Regarding the definition of a social alignment for a community, we have only looked at Stack Overflow in isolation. Like many other web services, there is a relationship between the platform and search engines, mainly Google [25]. Searching for the answer to a programming-related question on Google and being directed to Stack Overflow results in different posts being shown than when using Stack Overflow's own search engine. Figure 21 shows one example of such a difference regarding the query "python merge two lists".

The method a person uses to find posts on the website may have an effect on the community. A tag on Stack Overflow might be isolated from its broader community if the website is not ranked highly by Google's algorithm when considering related search terms. Such a community may show differences in its users behavior and transition.

As there are many more Stack Exchange websites, which are all very similar to Stack Overflow, and non-English versions of the platform, it might be worth to apply our methods to these websites, comparing the results. Other platforms may show different transition behavior because of the topics covered there and possible differences in the users on the various platforms regarding individual features. Non-English Stack Overflow variants may be used, for example, to compare transition for individuals active on both the main and more specific website. This might help in finding out if the specific platforms actually help in overcoming barriers.

References

[1] Konstantinos Adamidis. "Theory & Methods: An EM algorithm for estimating negative binomial parameters". In: *Australian & New Zealand journal of statistics* 41.2 (1999), pp. 213–221.

[2] Adelin Albert and John A Anderson. "On the existence of maximum likelihood estimates in logistic regression models". In: *Biometrika* 71.1 (1984), pp. 1–10.

[3] Miltiadis Allamanis and Charles Sutton. "Why, when, and what: analyzing stack overflow questions by topic, type, and code". In: *2013 10th Working Conference on Mining Software Repositories (MSR)*. IEEE. 2013, pp. 53–56.

[4] Kristen M Altenburger et al. “Are there gender differences in professional self-promotion? an empirical case study of linkedin profiles among recent mba graduates”. In: *Eleventh International AAAI Conference on Web and Social Media*. 2017.

[5] Sebastian Baltes et al. “Sotorrent: Reconstructing and analyzing the evolution of stack overflow posts”. In: *Proceedings of the 15th international conference on mining software repositories*. 2018, pp. 319–330.

[6] Amiangshu Bosu et al. “Building reputation in stackoverflow: an empirical investigation”. In: *2013 10th Working Conference on Mining Software Repositories (MSR)*. IEEE. 2013, pp. 89–92.

[7] Miguel Botto-Tobar et al. “Is stack overflow in portuguese attractive for brazilian users?” In: *Proceedings of the 13th International Conference on Global Software Engineering*. 2018, pp. 21–29.

[8] Margaret Burnett et al. “GenderMag: A method for evaluating software’s gender inclusiveness”. In: *Interacting with Computers* 28.6 (2016), pp. 760–787.

[9] Wenhong Chen and Barry Wellman. “The global digital divide-within and between countries”. In: *It&Society* 1.7 (2004), pp. 39–45.

[10] Maëlick Claes et al. “Do programmers work at night or during the weekend?” In: *Proceedings of the 40th International Conference on Software Engineering*. 2018, pp. 705–715.

[11] Laura Dabbish et al. “Social coding in GitHub: transparency and collaboration in an open software repository”. In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM. 2012, pp. 1277–1286.

[12] Cristian Danescu-Niculescu-Mizil et al. “A computational approach to politeness with application to social factors”. In: *arXiv preprint arXiv:1306.6078* (2013).

[13] Cristian Danescu-Niculescu-Mizil et al. “No country for old members: User lifecycle and linguistic change in online communities”. In: *Proceedings of the 22nd international conference on World Wide Web*. 2013, pp. 307–318.

[14] Nadia Eghbal. *Roads and bridges: The unseen labor behind our digital infrastructure*. Ford Foundation, 2016.

[15] Thomas Erickson and Wendy A Kellogg. “Social translucence: an approach to designing systems that support social processes”. In: *ACM transactions on computer-human interaction (TOCHI)* 7.1 (2000), pp. 59–83.

[16] Denae Ford, Alisse Harkins, and Chris Parnin. “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)*. IEEE. 2017, pp. 239–243.

[17] Denae Ford et al. “Paradise unplugged: Identifying barriers for female participation on stack overflow”. In: *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM. 2016, pp. 846–857.

[18] Denae Ford et al. “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*. ACM. 2018, p. 608.

[19] Cristian González Garcıa et al. “Swift vs. objective-c: A new programming language”. In: *IJIMAI* 3.3 (2015), pp. 74–81.

[20] Philip J Guo. “Non-native english speakers learning computer programming: Barriers, desires, and design opportunities”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 396.

[21] James J Heckman. “Sample selection bias as a specification error”. In: *Econometrica: Journal of the econometric society* (1979), pp. 153–161.

[22] Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

[23] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[24] Anna May, Johannes Wachs, and Anikó Hannák. “Gender differences in participation and reward on Stack Overflow”. In: *Empirical Software Engineering* (2019), pp. 1–23.

[25] Connor McMahon, Isaac Johnson, and Brent Hecht. “The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies”. In: *Eleventh International AAAI Conference on Web and Social Media*. 2017.

[26] Dana Movshovitz-Attias et al. “Analysis of the reputation system and user contributions on a question answering website: Stackoverflow”. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM. 2013, pp. 886–893.

[27] Nigini Oliveira, Nazareno Andrade, and Katharina Reinecke. “Participation differences in Q&A sites across countries: opportunities for cultural adaptation”. In: *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. 2016, pp. 1–10.

[28] Helen Peterson. “The gendered construction of technical self-confidence: Women’s negotiated positions in male-dominated, technical work settings”. In: *International Journal of Gender, Science and Technology* 2.1 (2010).

[29] Marcel Rebouças et al. “An empirical study on the usage of the swift programming language”. In: *2016 IEEE 23rd international conference on software analysis, evolution, and reengineering (SANER)*. Vol. 1. IEEE. 2016, pp. 634–638.

[30] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. “Extracting the multiscale backbone of complex weighted networks”. In: *Proceedings of the national academy of sciences* 106.16 (2009), pp. 6483–6488.

[31] Aaron Shaw and Eszter Hargittai. “The pipeline of online participation inequalities: the case of wikipedia editing”. In: *Journal of Communication* 68.1 (2018), pp. 143–168.

[32] Dan Sholler et al. “Ten simple rules for helping newcomers become contributors to open projects”. In: *PLoS Computational Biology* 15.9 (2019).

[33] Rogier Slag, Mike de Waard, and Alberto Bacchelli. “One-day flies on stackoverflow-why the vast majority of stackoverflow users only posts once”. In: *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*. IEEE. 2015, pp. 458–461.

[34] Ivan Srba and Maria Bielikova. “Why is stack overflow failing? preserving sustainability in community question answering”. In: *IEEE Software* 33.4 (2016), pp. 80–89.

[35] Stack Overflow. *Developer Survey Results*. [Online; accessed 29-September-2019]. 2019. URL: https://insights.stackoverflow.com/survey/2019.

[36] Stack Overflow. *Non-English Question Policy*. [Online; accessed 08-January-2020]. 2009. URL: https://stackoverflow.blog/2009/07/23/non-english-question-policy/.

[37] Stack Overflow. *Stack Overflow Isn't Very Welcoming. It's Time for That to Change.* [Online; accessed 03-January-2020]. 2018. URL: https://stackoverflow.blog/2018/04/26/stack-overflow-isnt-very-welcoming-its-time-for-that-to-change/.

[38] Klaus Stein and Claudia Hess. “Does it matter who contributes: a study on featured articles in the german wikipedia”. In: *Proceedings of the eighteenth conference on Hypertext and hypermedia*. ACM. 2007, pp. 171–174.

[39] Igor Steinmacher et al. “A systematic literature review on the barriers faced by newcomers to open source software projects”. In: *Information and Software Technology* 59 (2015), pp. 67–85.

[40] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. “Stackoverflow and github: Associations between software development and crowdsourced knowledge”. In: *2013 International Conference on Social Computing*. IEEE. 2013, pp. 188–195.

[41] Claudia Wagner et al. “It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia”. In: *Ninth international AAAI conference on web and social media*. 2015.

[42] Xin Xia et al. “What do developers search for on the web?” In: *Empirical Software Engineering* 22.6 (2017), pp. 3149–3185.

[43] Lei Xu, Tingting Nian, and Luis Cabral. “What makes geeks tick? a study of stack overflow careers”. In: *Management Science* (2019).

[44] Jiang Yang et al. “Culture matters: A survey study of social Q&A behavior”. In: *Fifth International AAAI Conference on Weblogs and Social Media*. 2011.