A Tale of Two Virtual Communities: A comparative analysis of culture and discourse in two online programming communities

Subhasree Sengupta
School of Information Studies
Syracuse University
susengup@syr.edu

Abstract

Software programming is increasingly becoming a collaborative and community driven effort, with online discussions becoming vital resources for learning and knowledge sharing. This study explores the differences in the discourse patterns of two popular online programming communities and provides insights into the type of community practices and learning outcomes these collectives support and scaffold. A three step content analysis framework is presented that employs a mixture of automated text processing techniques and qualitative methods on a representative sample of 8639 and 6126 contributions from Stack Overflow and r/Askprogramming respectively. Results indicate differences between communities emerge in the scope of topics and the nature of responses the community provides. While r/Askprogramming has a more community centric, interpersonal approach and provides a space for sharing and supporting needs beyond knowledge sharing and factual learning, Stack Overflow takes a more task focused, knowledge centric approach. These findings suggest key normative structures that regulate patterns of collaboration and deliberation, which may have long term design implications for structuring and sustaining informal learning initiatives that nurture and promote technical skill development and enhancement.

1. Introduction

Education and learning are integral to the intellectual growth of our society. What we learn builds our character, our value systems, and shapes us into the type of individuals we become. In today's world, with the growing popularly of automation and data science skills, acquiring and having technical, particularly programming knowledge and expertise is becoming important and essential [1].

Learning to code relies on application skills, creative acumen, and tacit knowledge and experience. Hence, in software programming, learning through group discussion and deliberation beyond traditional modes of instruction holds of promise for continued acquisition of expertise [2]. In this regard, prior work has highlighted the effectiveness of learning through conversation and collaboration [3]. With the growth of the internet and advancements in communication infrastructure, online forums can help to boost connectivity and collaborative learning practices, which further motivates the potential to design virtual learning collectives that nurture learning technical skills through collaborative discourse. This research explores how interaction in online forums focused on computer programming, supports both learning and community development.

Two popular channels of discourse on programming and software development are Stack Overflow and r/Askprogramming. Created in 2008, Stack Overflow is a popular forum that serves over 50 million users worldwide, of which the vast majority are working professionals and students who wish to take up computing as a career. The subreddit r/Askprogramming is a forum that has similar usage as Stack Overflow. With over 48000 users, this subreddit supports deliberation on a wide variety of topics related to programming and software related careers. Although the user base is not as large as Stack Overflow, use of such Reddit channels for extending discussions on technical skill development is gaining momentum. This motivates the question of understanding how such virtual channels are used and the type of learning needs and community dynamics such collectives support.

This study extends and forms a part of a larger pool of work on learning through exploratory dialogue and collaboration in the context of online communities [4, 5, 6]. However, the type of learning can by greatly influenced by the culture, normative practices and organizational structure of such online communities. Thus, the goal of this work is to add further nuance to this larger body of work, by providing insight into the type of community structures and norms evident in the discourse patterns of Stack Overflow

URI: https://hdl.handle.net/10125/70616
978-0-9981331-4-0

and r/Askprogramming. By conducting a comparative analysis, we further explore the type of learning support and mentoring different communities might offer, which can help to provide insights into the question of design and structuration of online informal learning initiatives that support technical learning.

2. Background and theoretical motivation

2.1. Learning in virtual communities

This study is primarily grounded in the theories of social constructivism and social learning [7, 8], which highlight that learning is not only a cognitive process but also influenced by the social context of the learner. This further emphasizes the importance of the community in understanding the type of learning processes it supports. Communities in the online space represent a network of individuals connected through some shared interest that leads to the development of interpersonal bonds and a collective and shared community identity, which may impact the type of discussions and type of support provided [9, 10, 11]. The type of learning, nurturing and support a community provides can depend on the type of relationships and social capital fostered in the community [12]. Weak relationships or bridging capital may offer diversification of knowledge through connections with virtual strangers, whereas strong tie relationships or bonding capital may provide support, facilitate trust building and lead to a greater sense of attachment and affinity among community members [13, 14]. The nature of relationships a community supports, depends on the type of culture and commitment valued and practiced by its members. This further motivates the need to understand the norms, implicit organizational practices of online communities and the impact of such activities on the nature of discourse and learning observed in such spaces.

2.2. Learning and the impact of community structure

Discourse and community engagement can greatly depend on the interplay between community structure and individual agency. This interplay of structure and agency, termed as *dualism of structure* [15], can impact the patterns of repeated engagement, participation, establishment of virtual relationships and serve as the key force that defines and distinguishes the culture, discourse norms and the collective identity of a community [16, 17, 18]. Norms emerge out of these structural dualisms and represent informal rules that drive participation and conversation in online channels [19]. As mentioned in [17], norms are key contextual variables that drive feedback mechanisms, reputation building and perceived implicit hierarchies in online forums. Norms and community culture may impact the extent of expression [20, 21], nature of trust, intimacy and attachment [10], collaboration and distribution of tasks [22], all of which may affect the processes of peer driven knowledge production and discussion or the nature of *epistemic culture* [23] and *values* [24] promoted in these collectives.

2.3. Learning and the issue of design

Design of a virtual learning space is intricately tied with the issues of structuration and normative practices that drive the interaction patterns in online communities. Hence, this raises the question of how design affects the community dynamics, virtual relationships and social-cultural affordances associated with such communities and vice-versa [25]. Learning in online communities is driven by spontaneous or *over the shoulder learning* contributions based on personal experience and expertise [26]. Thus design in this context can be an essential in increasing the effectiveness and learning efficacy associated with these platforms. In this regard, [27] presents a two mode model that highlights two main patterns of contributions in online peer communities. Essentially, this framework uses the theme of weak vs strong tie patterns to establish two models of contributory behavior - (1) Lightweight, akin to weak tie models of contributions, which are more focused on knowledge development and are more task based and (2) Heavyweight, akin to strong tie based contributions, which encourage development of stronger intra-communal bonds and foster a sense of belongingness to the community. Prior work in this context has highlighted the importance of both models as each serves different functions and objectives in online spaces [28]. This motivates the context of exploration of this study, on understanding different contributory patterns that can be discerned from the conversations of the two communities, in order to address the larger question of design of such online learning venues.

3. Research questions Development

Language, culture, and community organization even in the virtual context are intricately tied together [29]. Since textual conversations are the primary mode of interaction in these virtual spaces, investigating the differences in discourse can help to provide a deeper understanding of community values, collective beliefs that manifest in the discourse. Using the theoretical framework lightweight and heavyweight models of contributions [27], this study is aimed at understanding the differences in the learning and community practices in the discourse of two virtual programming communities (Stack Overflow and r/Askprogramming). Although some prior research has addressed the nature of discourse and participation on Stack Overflow [21, 30], most studies have focused on single platforms or technologies, even when examining different groups with that platform. To address this gap, this study uses a comparative analysis approach to evaluate interaction in two popular programming learning forums. The overall research question is:

What kind of implicit community structures manifest in discourse patterns, and how do these structures relate to learning in open online forums?

The main research questions that this study aims to address in the context of Stack Overflow and r/Askprogramming are:

1) What kind of implicit community norms are indicated through the discourse?

2) How do these norms differ across the communities?

4. Analysis Method

Since textual conversation is the primary mode of collaboration and knowledge construction in these online discussion channels, the main motivation for the analysis was to explore and uncover latent topics embedded in the discourse of two programming forums (r/Askprogramming and Stack Overflow). A three step mixed methods approach was employed. The first step was an exploratory analysis of a small sample of 15 posts from both Stack Overflow and r/Askprogramming. This step provided an initial understanding of the type of topical themes discussed in each of these forums. The second step used automated content analysis approaches, such as topic modeling and topic coherence, to find initial estimates of topical themes embedded in the larger data set 1. The final step involved a qualitative investigation of the topic labels identified by the quantitative framework to find the final set of topic themes.

4.1. Data collection

Stack Overflow: Data from Stack Overflow was collected from a combination of publicly available APIs (https://stackapi.readthedocs.io/en/latest/) and web scraping scripts, implemented in Python. For this analysis we focus on the posts created between January and December 2019. However, the total number of contributions to Stack Overflow created in this time frame is around 5 Million, and analyzing the full corpus is beyond the scope of this analysis. A two step process was followed for extracting a sample. First, the python wrapper for the Stack API was used to gather a random sample of 500 questions for posts in Stack Overflow created between January to December, 2019. A Stack Overflow post consists of three main textual elements - (1) Questions (2) Answers (3) Comments. The second step used, a web scraping script to collect the answers and comments for each of the 500 questions, gathered in the first step. Altogether, the data sample has 8639 contributions, which included all questions, answers and comments.

Table 1. Data statistics for Stack Overflow(SO) and r/Askprogramming(r_Ask)

Statistics	**SO**	**r_Ask**
# of contributions	8639	6126
mean score	1.57	2.4
# of unique users	3172	2005
earliest creation date	01/02/2019	01/01/2019
latest creation date	11/28/2019	08/31/2019

r/Askprogramming: To maintain consistency, the data collection procedure for this site was set up to mirror that of Stack Overflow as much as possible. For collecting data from r/Askprogramming, a publicly available cloud based data repository, Google Bigquery (https://cloud.google.com/bigquery/) was used. Bigquery was used as it houses a large collection of Reddit posts, starting from 2005 to 2019. Since r/Askprogramming as a community is not as extensively used as Stack Overflow, in order to ensure that the total number of contributions were nearly comparable across both communities, a random set of 1000 questions was sampled from the r/Askprogramming corpus for the year 2019. Reddit posts are made of two primary types of contributions: (1) questions (2) comments. On Bigquery, questions and comments are stored in separate databases. Further, post data is segregated by month. A three step process was followed to collect the data sample. First, all question and comment data across all the months as stored on the bigquery data server was collected using a python connector to the database. Second, the questions were aggregated and a random sample of 1000 selected for analysis. Third, all comments were aggregated and those corresponding to the 1000 questions shortlisted in the previous step were selected for analysis. On completing this process, the data set included 6126 contributions including questions and comments.

4.2. Analysis pipeline:

Category derivation: The first layer of analysis entailed an exploratory investigation of the two communities. The goal of this analysis was to gain an understanding of the type and number of topics present in the discourse patterns of these two communities. In order to conduct this exploration, fifteen posts (with 236 contributions for r/Askprogramming and 224 for Stack Overflow) were selected, using the same data collection methods followed for gathering the larger sample (as described in section 4.1). A qualitative thematic coding procedure as outlined in [31] was used to analyze the data.

For each data set the following steps were used to derive the initial set of topical themes. First, for each post an aggregate score which was the sum of the actual score (calculated as the absolute value difference of the number of upvotes and downvotes) the post received and the total number of contributions associated with the post was computed. Further, the posts were sorted based on this aggregate score. Next, all the contributions (questions, answers, comments) associated with the top five posts (based on the aggregate scores determined in the first step) were analyzed to find the topical themes embedded in the contributions. Once a set of themes were identified, the remaining contributions were coded according to these themes. Modifications and refinements were made to the initial coding schema as applicable based on the remaining set of contributions. Finally, this coding schema was validated across all the contributions of the fifteen posts to ensure that no new themes emerged.

This coding process was followed for both Stack Overflow and r/Askprogramming datasets and a separate set of themes derived for each, which were then compared and contrasted. At the end of this initial analysis an initial coding schema with a set of 14 and 16 topical themes emerged for Stack Overflow and r/Askprogramming, respectively.

Automated processing: This step forms the core of the analysis pipeline. The goal of this layer of analysis was twofold. The first aim was to use computational methods to scale the analysis to a larger representative sample. The second goal was to address some of the challenges that may exist in traditional content analysis techniques. Content analysis involves segregating data into topical clusters which can be a challenging process, when done purely based on human input. Computational methods are particularly helpful in this regard. By leveraging syntactical and semantic properties of the data, these techniques can help in providing insights into how data can be partitioned into different non-overlapping themes. In this context, topic modeling [32] is a popular computational approach that is used extensively to automatically detect topical clusters.

Topic-modeling is an unsupervised machine learning approach useful for finding latent topics distributed across the data. Topic modeling leverages statistical models to detect hidden semantic structures in order to automatically cluster the data. Essentially it uses co-occurrences of words in a data set to find patterns or themes (latent topics), based on which the data can be partitioned. The non-negative matrix factorization (NMF) approach [33] was used for topic modeling since it has been found to be more robust and scalable than other approaches [34]. To make the NMF analysis more robust additional preprocessing steps such as stop-word removal and normalization of the document term matrix were also carried out [34]. NMF was applied to both Stack Overflow and r/Askprogramming datasets separately. The NMF models and preprocessing steps were implemented using the scikit-learn machine learning library in Python [35].

A drawback of using topic modeling is the need to have an estimate of the number of topics prior to running the model. Since this is an unsupervised method, it is hard to have an estimate of the exact number of topics that best represents the topic distribution of the data corpus [34, 36]. In this context, topic coherence is a technique that can be used to get an estimate of the number of topics. The main idea behind topic coherence is to create a quantitative approach to evaluate how accurately topical clusters describe the main semantic structures embedded in a data set [37, 34]. To do so, each topic model extracted from the data corpus is assigned a *coherence score*, which is a measure of how semantically close words in a topic are to one another. Word2Vec [38] is a popular and robust method for computing semantic similarity between word pairs and is frequently used to compute coherence scores for topic models [34]. The coherence scores and Word2Vec models were implemented using the gensim library in python [39] for this study.

For this study, an end-to-end automated processing approach that combined topic coherence with topic modeling was used. In doing so, the problem of estimating the number of topics associated with topic modeling was addressed. The steps followed for this process were as follows. First a range of topics (k_{min}, k_{max}) was determined, with k_{min} set to 5 and k_{max} set to 25. This range was selected such that the mean k was 15, which was also the average number of topics determined in the initial exploration across Stack Overflow and r/Askprogramming. For each topic k, an

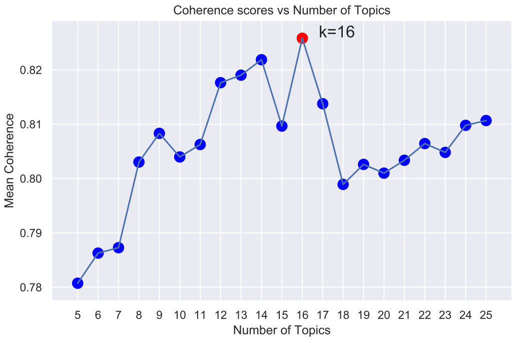

Figure 1. Plot of coherence scores for R/AskProgramming

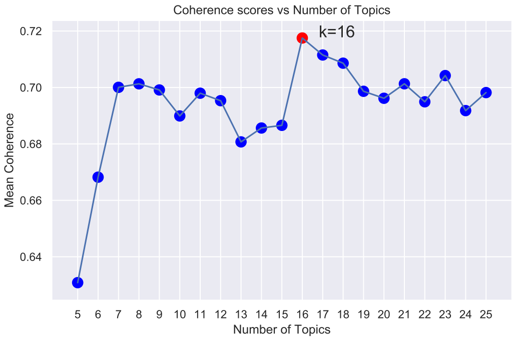

Figure 2. Plot of coherence scores for Stackoverflow

NMF model was created and the coherence score for this model for determined. Finally, the coherence scores for all the models were compared and the model which had the maximum coherence score was selected as the final model. Figures 1, 2 represent the coherence scores for r/Askprogramming and Stack Overflow respectively. For both Stack Overflow and r/Askprogramming, the NMF model with 16 topics has the best topical estimates for the data corpus. In the following step, these results were further pruned and refined.

Qualitative refinement: The final and most crucial step this analysis pipeline, was aimed at combining and refining the inferences of the previous steps, while also ensuring that the nuances of the results were maintained. The main objective of this step was the label and add contextual meaning to the topics determined in the automated processing step, using the preliminary coding schema developed in the initial category derivation step. To infer the topics, the following steps were followed; first the data set was sorted based on score, and then for each topic the top 25% of the contributions were selected and coded, in order to validate the initial findings. The decision was made since on manual inspection it was found that the top 25% of the data associated with each topic, best represented the semantic meaning of that topic label. At the end of this stage, a final coding schema of 14 topic themes (across both communities) was ascertained which shall be described in the subsequent section.

5. Emergent Content categories

A total of 10 themes from Stack Overflow and 14 themes from r/Askprogramming were determined based on the final qualitative refinement, done in the last step of the analysis pipeline. 2 presents the percent of responses for each of the 14 themes across Stack Overflow and r/Askprogramming. These themes can further be grouped into three broad perspectives - (1) Knowledge perspective (2) Community perspective (3) Combined perspective. These are elaborated next.

5.1. Knowledge perspective

These contributions are more knowledge centric, aimed at providing professional enrichment and mentoring [40]. Akin to the idea of direct collaboration [41], these help to provide direct and fast resolution to questions. Similar to the idea of Lightweight models of knowledge production [27], these contributions demonstrate a more individualistic sense of attachment and commitment [24] and thus are instances of the bridging capital [12] created within the community.

Solution strategy: Aimed at providing direct solutions, these contributions usually had a sequence of steps with some explanation needed to resolve the problems stated in prior questions. An example quote for this category stated *Store the rotation in a data attribute, increase it on each click and set the style basing on that value*, which was given as a response to a question that asked about implementing a feature using a web programming language on Stack Overflow. This type of contributions form a significant majority of Stack Overflow contributions as compared to r/Askprogramming.

Solutions with only code: A subset of contributions more frequently found on Stack Overflow provided quick and prompt solutions in the form of code snippets that can be directly applied to resolve the problem presented in prior contributions. Such contributions, although more knowledge oriented demonstrate immediateness and promptitude among community members to provide quick corrections to errors.

Debugging: These contributions were questions posters asked in the both communities about correcting code and involved learning syntax, logical flow or understanding how to apply, expand existing solutions. This was the most common theme of questions around which several discussions were centered. These type of contributions were more frequently found in the interactions of Stack Overflow than that of r/Askprogramming.

Software Ideation: A subset of contributions on r/Askprogramming pitched and enquired about tools, applications that can be created using certain software, programming languages. An example quote in r/Askprogramming for this category stated *I have an idea for a private message board to be used by a private group. Does any know of an open source project that I can use or provide further tips?*. These contributions highlight how these communities support idea and creative development, help to nurture talent and provide professional enrichment for community members.

5.2. Community perspective

These contributions were aimed at fostering strong interpersonal relationships among community members and represent the bonding capital created among community members [12]. Akin to the findings of [42, 43], these contributions help provide emotional support, solidarity and help to build resilience among community members. Similar to the idea of heavyweight models of contributory behavior [27], these models depict a greater sense of attachment, trust and commitment to the collective interest of the community [24, 10, 44].

Programming humor: These contributions indicated socializing and developing rapport with others in the community. These discussions although centered on programming but were usually more humorous and were aimed at building camaraderie among community members. An example contribution in r/Askprogramming stated *In svg manipulation there s something called curveto and moveto I have heard these pronounced curvetto and movetto as if they were Italian music jargon*. Such contributions were only found in the r/Askprogramming community.

Navigating workplace challenges: A subset of discussions on r/Askprogramming involved asking and providing guidance regarding navigating toxic work situations and practices, managing team dynamics and improving relationships with co-workers. An example quote on r/Askprogramming for this type mentioned *I don't mean to be terse but I'm a development manager, its hard to constantly motivate employees, there is lack of innovation, initiative, its difficult to find a metric to evaluate and reward human endeavor*. This contribution sought suggestions from community members about managing team members and evaluating work quality in order to improve management practices in organizations that engage software professionals. These findings reflect ease in disclosing moral dilemmas and professional struggles which indicates a strong sense of affinity, trust and understanding among community members who might be strangers to one another beyond the virtual sphere of communication [11, 13].

Providing encouragement: These contributions exemplify feelings of compassion and a commitment to encourage posters who disclose professional hurdles or personal issues related to learning software programming. A significant portion of these post provided instances of personal anecdotes, which indicate a greater commitment to the overall well being of the community [45]. An example quote stated *Its the hard part of corporate culture, what is most disturbing is that you were simply handed a list of to-dos to implement, it certainly would be hard for anyone*. These contributions were present only in r/Askprogramming.

Strategies for self development: These contributions were aimed at seeking and providing suggestions regarding enhancing career outcomes by boosting oneś self confidence and overcoming learning difficulties. An example quote in r/Askprogramming recommended using online coding tools to improve programming efficiency stated *I recently started using online coding for practice purposes, I spend 2-3 hours daily and solve just 2-3 problem , it really helps to improve my overall thinking and efficiency*. These not only demonstrate a dedication to the success of community members but also demonstrate the value of knowledge held within the community [40]. These contributions were present only in r/Askprogramming.

Validation: These contributions involve expressing gratitude, supporting and encouraging users for their thoughts and contributions. These demonstrate a normative convention of these communities to acknowledge and maintain a sense of decorum and politeness while conversing with one another [46]. An example from r/Askprogramming for this type of contribution stated *Thanks your explanation was very illuminating*. These contributions were more prominent in r/Askprogramming, but also present in Stack Overflow.

Community issues: These contributions involved discussions around applicability of a post, conventions to follow when contributing to the community. An example quote on r/Askprogramming explaining the conventions associated with flagging a post stated *Once you mark a question a duplicate you have to also shut the original question down for being a duplicate* *of the new question.* These depict a deep sense of understanding of the community and were aimed at improving alignment of discussions to the overall community objective and values [24]. Although present in small percentages across both communities, these were more prominent on r/Askprogramming.

5.3. Combined perspective

This type of contributions critique, clarify, expand and most importantly repair (restructuring to ensure alignment with community objectives) contributions especially questions such that the discussions are pertinent to the community [41, 47, 48] and are hence prominent in the conversations of both Stack Overflow and r/Askprogramming. These require both an understanding of the community values and also knowledge of the domain (e.g. programming) and thus represent a middle ground between lightweight and heavyweight models of contributions [27].

Clarification: These typically include explanations of concepts, rephrasing or modifications to prior contributions in a post in order to enhance the clarity or make them more applicable to both the community and the context of discussion. An example of this category on r/Askprogramming stated *sorry but do you mean the same computer or is it on something else?*, which is asking for a clarification regarding a question posted about a computing software. These posts were more prominent in the discussions on Stack Overflow as compared to r/Askprogramming.

Alternatives: These contributions are aimed at broadening the scope of previous contributions by augmenting further dimensions of thought, and can thus help to extend the context of application of the information shared in the community [41, 49]. An example of this category on r/Askprogramming was *Maybe you need to rethink your interfaces Sometimes long walls of code can cause this issue for me too and is usually a sign that maybe I should be abstracting things out more*. While this does not give the direct solution, but suggests an alternative to critically analyze their logical approach to the problem. These posts were more prominent in the discussions on Stack Overflow as compared to r/Askprogramming.

Limitations: These types of contributions are used to express shortcomings, lack of applicability in certain scenarios of previous contributions (especially solutions or clarifications). An example of this type of conversation on Stack Overflow stated as*Thanks. But that has thrown a new error.* , which highlights a flaw in the logical approach presented in a previous contribution. These posts were more prominent in the discussions on Stack Overflow as compared to r/Askprogramming.

Table 2. Percentage of responses in the categories across (SO) and (r_Ask)

Content category	% in SO	% in r_Ask
Solution strategy	24	4
Clarification	20	6
Alternatives	12	7
Limitations	10	2
Programming humor	4	10
Providing encouragement	0	6
Navigating workplace challenges	0	22
Strategies for self development	0	7
References	6	11
Debugging	9	5
Software ideation	0	5
Validation	5	11
Community issues	2	3
Solutions with only code	8	1

References: These contributions were aimed providing links to other discussions that discussed similar issues indicated in previous contributions. An example quote was *Have you read this post ...[URL of a similar post on Stack Overflow]* These demonstrate a deeper understanding the helpfulness of the community, that is an understanding of knowledge management practices and the quality of the information curated by the community [50, 51]. These posts were more prominent in the discussions on r/Askprogramming as compared to Stack Overflow.

6. Discussion and conclusion

The 14 topical themes that emerge in the comparative analysis highlight the different group processes, knowledge co-construction and curation that take place in these online forums. Further, clustered into the three perspectives of knowledge, community and the combined view highlights the values, norms and collective identity each of these forums hold [17].

While contributions around the knowledge perspective were more prominent in Stack Overflow, the community perspective was more evident in the communication traces of r/Askprogramming. The more interpersonal approach to conversations in the r/Askprogramming community indicates that this learning space in addition to providing instrumental support, mentoring and nurturing can also serve as a space to seek social and emotional support [43, 44]. The combined perspective is of particular interest, as it highlights a third dimension of contributory behavior, that is in between lightweight (knowledge perspective) and heavyweight models (community perspective) of contributions [27]. This might indicate that these models of knowledge production might manifest as a spectrum instead of a binary categorization. Similar to the idea of the *dualisms of structure* [15], this may further provide insights into how these communities evolve in their normative practices, in an effort to maintain alignment with the implicit organizational structure of the community but also provide agency to community members to express themselves as they may wish to.

The central issue in designing and structuring of such online learning spaces is to understand the type of environment that should be created, the level of intimacy or social openness [24] that should be supported in addition to knowledge and task based support. As indicated in [42], these online channels may help to create a *safe space* for individuals to disclose and seek guidance in navigating personal struggles, disclose sensitive issues (e.g., mental health challenges, work hazards). By sharing lessons learned through experiences, such collectives may provide guidance and encouragement in ways which may be absent or hard to find in more institutional learning environments. In this regard the themes of programming humor, providing emotional support, and suggestions for navigating workplace challenges, can become crucial dimensions of support such online environments may provide. Further, themes such as software ideation and strategies for self-development may indicate how these communities help to provide strategies to improve oneś approach and outlook towards professional activities, which can prove to be very helpful in their long term career growth. Having the opportunity for such expression can be essential in academic or learning environments and can help to develop a more equivocal, open and fair community that offers a multifaceted perspective towards learning and mentoring.

The most important issue of consideration in the issue of design and structuration is the type of values these communities support [52]. Through the comparative analysis, we see distinctive characteristics, norms or cultures that manifest in these spaces that drive the learning practices observed. While Stack Overflow maintains a knowledge centric community dynamic, r/Askprogramming fosters a culture that encourages the development of interpersonal bonds and a sense of belongingness among community members. The combined perspective, found in both Stack Overflow and r/Askprogramming represents both a knowledge and a more social or communal dynamic in the community. Such contributions indicate that a subset of community members not only address the information needs of the community but also consider the goals and objectives of the community when contributing to these collectives. Such contributions become very important when considering the type of social capital these online channels build and the awareness that community members have towards the knowledge management practices of these collectives [10].

Thus these insights indicate that such online venues may hold potential to serve additional needs or be important for augmenting support that traditional or more formalized pedagogical systems may not be able to provide [26]. The three perspectives and differences in the community dynamics highlight the different learning or supportive needs that may be crucial in designing the functions of these online learning support initiatives. The insights, findings of this study can be helpful for improving moderation or team building activities that drive the discussions in these spaces. The knowledge of the different dimensions of support we indicate in this study can help to regulate the discourse and the scope of discussions in these communities. Further, these may also be essential to understand factors that affect how people collaborate and deliberate together in online collective environments. Recently, automation has become an essential aspect of online platforms, examples include the infusion of recommendation systems, team building software, and automated content curation (e.g., moderation) all of which are aimed at improving the management and efficacy of these online initiatives [53, 54]. A deeper and nuanced understanding of the community needs, values and nature of participation based on the insights gleaned from this study can help designers and practitioners to make an informed decision about how automated functionality should be augmented and how these should engage and collaborate with users, in order to create conducive online informal learning collectives that support and nurture technical and programming skill development.

7. Future work

Future work will include expanding and refining the content schema developed by extending the analysis to a larger data corpus and by further refining the analysis framework presented in this study. Additional dimensions of linguistic features of the discourse will be explored using natural language tool kits in order to further consolidate the findings. To further understand the nature of contributory behavior, particularly the spectrum of expression indicated in this study; the goal of future work will be more deeply engage with community members and users to understand the factors affecting participation and engagement in these communities. Such explorations will also help to further unpack the interplay between agency and structure in these communities. In future studies, a survey instrument will be developed and interviews with community members will be conducted in order to get a holistic understanding of the forces that drive the group dynamics and help sustain participation in these communities.

References

[1] A. Balanskat and K. Engelhardt, *Computing our future: Computer programming and coding-Priorities, school curricula and initiatives across Europe*. European Schoolnet, 2014.

[2] S. Popat and L. Starkey, "Learning to code or coding to learn? a systematic review," *Computers & Education*, vol. 128, pp. 365–376, 2019.

[3] D. Laurillard, *Rethinking university teaching: A conversational framework for the effective use of learning technologies*. Routledge, 2013.

[4] S. B. Shum and R. Ferguson, "Social learning analytics," *Journal of educational technology & society*, vol. 15, no. 3, pp. 3–26, 2012.

[5] S. R. Hiltz, "Collaborative learning in asynchronous learning networks: Building learning communities.," 1998.

[6] C. Haythornthwaite, P. Kumar, A. Gruzd, S. Gilbert, M. Esteve del Valle, and D. Paulin, "Learning in the wild: coding for learning and practice on reddit," *Learning, Media and Technology*, vol. 43, no. 3, pp. 219–235, 2018.

[7] L. S. Vygotsky, *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.

[8] A. Bandura and R. H. Walters, *Social learning theory*, vol. 1. Prentice-hall Englewood Cliffs, NJ, 1977.

[9] C. Haythornthwaite, "Social networks and online community," *The Oxford handbook of Internet psychology*, pp. 121–137, 2007.

[10] Y. Ren, F. M. Harper, S. Drenner, L. Terveen, S. Kiesler, J. Riedl, and R. E. Kraut, "Building member attachment in online communities: Applying theories of group identity and interpersonal bonds," *Mis Quarterly*, pp. 841–864, 2012.

[11] M. Shelton, K. Lo, and B. Nardi, "Online media forums as separate social lives: A qualitative study of disclosure within and beyond reddit," *iConference 2015 Proceedings*, 2015.

[12] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of facebook "friends:" social capital and college students' use of online social network sites," *Journal of computer-mediated communication*, vol. 12, no. 4, pp. 1143–1168, 2007.

[13] D. Constant, L. Sproull, and S. Kiesler, "The kindness of strangers: The usefulness of electronic weak ties for technical advice," *Organization science*, vol. 7, no. 2, pp. 119–135, 1996.

[14] P. Norris, "The bridging and bonding role of online communities," 2002.

[15] A. Giddens, *The constitution of society: Outline of the theory of structuration*. Univ of California Press, 1984.

[16] J. Godara, P. Isenhour, and A. Kavanaugh, "The efficacy of knowledge sharing in centralized and self-organizing online communities: Weblog networks vs. discussion forums," in *2009 42nd Hawaii International Conference on System Sciences*, pp. 1–10, IEEE, 2009.

[17] H. Rosenbaum and P. Shachaf, "A structuration approach to online communities of practice: The case of q&a communities," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 9, pp. 1933–1944, 2010.

[18] E. Wenger, "Communities of practice: A brief introduction," 2011.

[19] R. M. Martey, J. Stromer-Galley, J. Banks, J. Wu, and M. Consalvo, "The strategic female: gender-switching and player behavior in online games," *Information, Communication & Society*, vol. 17, no. 3, pp. 286–300, 2014.

[20] K. Crowston and I. Fagnot, "Stages of motivation for contributing user-generated content: A theory and empirical test," *International Journal of Human-Computer Studies*, vol. 109, pp. 89–101, 2018.

[21] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: a case study of stack overflow," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 850–858, 2012.

[22] A. Kittur, B. Lee, and R. E. Kraut, "Coordination in collective intelligence: the role of team structure and task interdependence," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1495–1504, 2009.

[23] K. K. Cetina, *Epistemic cultures: How the sciences make knowledge*. Harvard University Press, 2009.

[24] N. Oliveira, M. Muller, N. Andrade, and K. Reinecke, "The exchange in stackexchange: Divergences between stack overflow and its culturally diverse participants," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–22, 2018.

[25] D. Vyas, C. M. Chisalita, and G. C. Van Der Veer, "Affordance in interaction," in *Proceedings of the 13th Eurpoean conference on Cognitive ergonomics: trust and control in complex socio-technical systems*, pp. 92–99, 2006.

[26] M. B. Twidale, "Over the shoulder learning: supporting brief informal learning," *Computer supported cooperative work (CSCW)*, vol. 14, no. 6, pp. 505–547, 2005.

[27] C. Haythornthwaite, "Crowds and communities: Light and heavyweight models of peer production," in *2009 42nd Hawaii international conference on system sciences*, pp. 1–10, IEEE, 2009.

[28] N. R. Budhathoki and C. Haythornthwaite, "Motivation for open collaboration: Crowd and community models and the case of openstreetmap," *American Behavioral Scientist*, vol. 57, no. 5, pp. 548–575, 2013.

[29] L. Cherny *et al.*, "Conversation and community: Chat in a virtual world," tech. rep., 1999.

[30] S. Sengupta and C. Haythornthwaite, "Learning with comments: An analysis of comments and community on stack overflow," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[31] R. E. Boyatzis, *Transforming qualitative information: Thematic analysis and code development*. sage, 1998.

[32] D. M. Blei and J. D. Lafferty, "Topic models," *Text mining: classification, clustering, and applications*, vol. 10, no. 71, p. 34, 2009.

[33] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, pp. 556–562, 2001.

[34] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[36] M. Belford, B. Mac Namee, and D. Greene, "Stability of topic modeling via matrix factorization," *Expert Systems with Applications*, vol. 91, pp. 159–169, 2018.

[37] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 100–108, 2010.

[38] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[39] R. Řehřek and P. Sojka, "Gensim—statistical semantics in python," *Retrieved from genism. org*, 2011.

[40] E. Kariri and C. Rodríguez, "e-mentoring activities in online programming communities: An empirical study on stack overflow," in *Service Research and Innovation*, pp. 123–138, Springer, 2018.

[41] Y. R. Tausczik, A. Kittur, and R. E. Kraut, "Collaborative problem solving: A study of mathoverflow," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 355–367, 2014.

[42] S. Sengupta, "What are academic subreddits talking about? a comparative analysis of r/academia and r/gradschool," in *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pp. 357–361, 2019.

[43] N. Andalibi, P. Ozturk, and A. Forte, "Sensitive self-disclosures, responses, and social support on instagram: the case of# depression," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pp. 1485–1500, 2017.

[44] M. De Choudhury and S. De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[45] D. Y. Wohn and C. Lampe, "Psychological wellbeing as an explanation of user engagement in the lifecycle of online community participation," in *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, pp. 184–195, 2018.

[46] D. Williams, T. L. Kennedy, and R. J. Moore, "Behind the avatar: The patterns, practices, and functions of role playing in mmos," *Games and Culture*, vol. 6, no. 2, pp. 171–200, 2011.

[47] J. Meredith and E. Stokoe, "Repair: Comparing facebook 'chat'with spoken interaction," *Discourse & communication*, vol. 8, no. 2, pp. 181–207, 2014.

[48] A. W. Vargo and S. Matsubara, "Corrective or critical? commenting on bad questions in q&a," *IConference 2016 Proceedings*, 2016.

[49] R. Ferguson, Z. Wei, Y. He, and S. Buckingham Shum, "An evaluation of learning analytics to identify exploratory dialogue in online discussions," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 85–93, 2013.

[50] K. Chalkiti and M. Sigala, "Information sharing and knowledge creation in online forums: The case of the greek online forum 'dialogoi'," *Current Issues in Tourism*, vol. 11, no. 5, pp. 381–406, 2008.

[51] J. Otterbacher, "'helpfulness' in online communities: a measure of message quality," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 955–964, 2009.

[52] S. Costanza-Chock, "Design justice: towards an intersectional feminist framework for design theory and practice," *Proceedings of the Design Research Society*, 2018.

[53] R. W. Gehl and M. Bakardjieva, *Socialbots and their friends: Digital media and the automation of sociality*. Taylor & Francis, 2016.

[54] E. M. Hastings, F. Jahanbakhsh, K. Karahalios, D. Marinov, and B. P. Bailey, "Structure or nurture? the effects of team-building activities and team composition on team outcomes," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–21, 2018.