Communities and Clusters: User Interactions in an Online Discussion Forum for Computer Science Education

Sukanya Kannan Moudgalya
Michigan State University, United States
moudgal1@msu.edu

K. Bret Staudt Willet
Michigan State University, United States
staudtwi@msu.edu

Abstract: Prior research has shown that teacher communities are important for social, cognitive, and affective growth of teachers. Further, teacher isolation often encourages teachers to join online teacher communities. In the U.S., computer science (CS) teachers are often isolated as the only educators teaching CS in their schools. Thus, identifying and studying various teacher communities that these CS educators can join is extremely important for teacher educators; Computer Science Educators Stack Exchange (CSEd SE) is an online discussion forum for those interested in CS Education. In this paper, CSEd SE was examined using principles of social network analysis. CSEd SE exhibited several favorable traits of a community, such as democratic representation of users, discussions around a multitude of topics, and presence of moderators and reciprocal conversations. These results indicate that well-designed online discussion forums could be apt venues for CS teachers to form communities and enhance their professional learning networks.

Introduction

The Stack Exchange Network is one of the largest online question-and-answer platforms in the world, founded in 2009 and boasting over 100 million unique monthly visitors by 2015 (Stack Exchange, n.d.). It has numerous subject-specific discussion forums for topics such as coding, mathematics, engineering, and photography. Stack Exchange has recently opened discussion forums for math educators (2013) and computer science (CS) educators (2016) as well. In this paper, we focus on users of the discussion forum Computer Science Educators Stack Exchange (CSEd SE), self-described as: "a question and answer site for those involved in the field of teaching Computer Science" (Computer Science Educators Stack Exchange, n.d.).

In this forum, registered users can ask questions on a wide range of topics related to CS education. Once they ask a question, they can assign it "tags," so that similar groups of questions can be identified by a common theme or idea. Examples of tags include "curriculum design," "student motivation," and "algorithm." Users can also *reply* to (i.e., answer) questions posed by other users, *comment* on both questions and answers, *upvote* and *downvote* questions and answers (i.e., show appreciation by giving a post a +1 score, or show displeasure by giving a -1 score), and so on. Through the voting mechanism, users are held accountable to what they post on the forum; the content on the forum is thus regulated by forum users and members to a large extent. The forum also has moderators and editors who have the power to edit posts, flag posts as being inappropriate, or 'lock' posts so that no one else can upvote or downvote the posts.

The users on CSEd SE included K-12 and college-level CS educators, working professionals in the software industry, students, and other people interested in CS education. Users also come from many different countries, such as the U.S., U.K., Canada, Ireland, India, Finland, etc. Each user has "reputation points" primarily based on the quality of their posts, such as user-generated votes their posts receive. As reputation scores ebb and flow across time, the changes are accessible for anyone to see.

Insights into the various interactions on the CSEd SE forum will help bring understanding to how participants tend to ask questions or answer them, whether or not users' reputation scores play a role, the extent to which questions and answers are upvoted, the kinds of topic tags that are popular on the forum, and if users' professions play a role in how they interact on the forum. Such insights could help researchers and practitioners identify and predict users who tend to give highly voted answers, start highly popular discussion threads, or have

high social capital in discussion forums catered towards teachers and educators. These insights could also help practitioners understand the kind of topics most popular on the forum, which would be beneficial for informing the kinds of topics that might be needed for teacher professional development programs.

Background

In the U.S., many K-12 CS teachers are isolated as the only CS educators in their schools; they are often referred to as "singletons" (Yadav, Gretter, Hambrusch, & Sands, 2017). This isolation impacts the opportunities for subject-specific peer-learning that these CS teachers might need. Teachers who feel isolated often seek online communities for peer-learning, growth, and development (Hur & Brush, 2009). Indeed, teacher communities and "professional learning networks" have been extremely successful in the past for teachers' social, cognitive, affective, and identity growth (Trust, Krutka, & Carpenter, 2016). In particular, there have been cases where high school CS teacher networks were central for inculcating a sense of community, promoting teacher reflection, and helping create a change in teaching practices (Ni, Guzdial, Tew, Morrison, & Galanos, 2011). As many CS teachers have time constraints, busy schedules, and geographical limitations to form offline CS communities, it becomes important to also foster online communities for CS educators (Ryoo, Goode, & Margolis, 2015). It is no surprise that online networks for CS teachers have been growing in recent years (Brown & Kölling, 2013).

A recent literature review of informal online teacher communities showed that the *communities of practice* framework has often been used to study teacher interactions in online forums (Macià & García, 2016). Communities of practice are defined as: "...groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis" (Wenger, McDermott, & Snyder, 2002, p. 4). These communities have the following traits: (a) a domain of shared interest; (b) people who engage in joint activities and discussions; and (c) members who are practitioners, such as educators. (Wenger, 1998b) Further, according to Wenger, they should also demonstrate (a) mutual engagement, such as dialogue instead of one directional information flow; (b) a joint enterprise; and (c) a shared repertoire. These three traits represent "...three dimensions of the relation by which practice is the source of coherence of a community" (Wenger, 1998a, p. 72). The mutual engagement component is especially important because one-directional sharing of information will not be representative of a community.

Given these characteristics of the community of practice framework, in this paper we explored traits of users who are *highly connected* on CSEd SE, which we define as those who have had at least three interactions with fellow users in terms of asking or answering questions—in other words, users who have high *degree centrality* (Kadushin, 2012). We chose three interactions as a baseline because this focused analysis on users who had interacted more than a back-and-forth exchange (i.e., two interactions)—evidence of CSEd as a community of practice would reside in users participating at least this much. We defined all the interactions between these highly connected users as CSEd SE's *core network of interactions*. In addition, more highly connected users might have high popularity or reputation and thus act as a source of social capital. To observe some the characteristics of highly connected users, we analyzed characteristics of the top ten most connected users—the top ten were chosen due to practical considerations related to the scope of this paper.

Previous studies on the Stack Exchange Network—in particular, the Stack Overflow website—have found that users with higher reputation scores tend to ask more questions and answer more highly voted questions (Movshovitz-Attias, Movshovitz-Attias, Steenkiste, & Faloutsos, 2013). Further, Movshovitz-Attias and colleagues found that the activities of users in the initial stages of joining the forum have been proven to be predictive of their future activities. In other words, it is possible to identify experts on these forums from the time they join a forum; this identification of expertise might be beneficial to practitioners in fostering and designing forums for educators. We also examined if there are *clusters* of users in the forum—that is, users who interact more with members in their cluster than they do with those outside it (Kadushin, 2012). Within clusters, there may be popular users or topics that help define the cluster—that is, users who might be connected with many people within the cluster due to their high levels of interactions and topics that get discussed most often within a cluster. Knowledge of this could help practitioners direct other educators interested in different topics to certain places or users on the forum.

With this background in mind, the following research questions guided this study:

1. What are the characteristics of users who contributed to the CSEd SE's core network of interactions?
2. What are the characteristics of the top-10 most connected contributors to the CSEd SE?
3. Does network clustering occur in the context of the CSEd SE's core network of interactions? If so, what are some characteristics of these clusters?

Method

Data collection

We collected data from the CSEd SE forum through data mining, using the statistical software *R* (R Core Team, 2018); specifically, we used the R package *stackr* (Robinson, 2015) to extract data from the CSEd SE website. The stackr package helps obtain the "read-only features of the Stack Exchange API with the ability to download information on questions, answers, users, tags, and other aspects of the site so that they can be analyzed in R" (Robinson, 2015).

Using stackr, we were able to gather metadata for CSEd SE questions, answers, comments, tags, and users from when CSEd SE started, May 2017, through June 2018. These metadata included information such as question identity number, answer identity number, user identity number, tag name, reputation scores of users, the date users joined the forum, and so on. This data collection resulted in a corpus of 559 questions, 2,675 answers, and 207 question tags from CSEd SE.

Data analyses

Once we obtained raw data in the form of questions, answers, comments, tags, and users' information, we reshaped the data for analysis in R—that is, we converted the data into a form that was suitable for further data analyses. One example of such data shaping was converting the "date-time" variable of when users joined CSEd SE into duration of forum membership. We then used the R package *igraph* (Csárdi, 2018) to create a graph of CSEd SE's core network of interactions, meaning we focused on *nodes* (i.e., contributors) in the network with degree centrality of three or higher. This helped us gather characteristics of the users who contributed to the CSEd SE's core network of interactions. Once we had a list of users who were the top contributors, we selected the top 10 and conducted content analysis by:

1. visiting the the top-10 users profile pages in the CSEd SE website,
2. reviewing each users' profile, and
3. conducting qualitative open coding to find themes or similarities in the users' profession, membership duration in CSEd SE, reputation points, location and relationship with fellow most-connected users.

Next, to examine whether or not clustering occurs in the core network of interactions, we used the *spinglass clustering algorithm* (Reichardt & Bornholdt, 2008). The spinglass clustering algorithm partitioned nodes into clusters by optimizing the function:

This function penalizes missing edges (i.e., non-links) of nodes (i.e., users) in the same cluster and links present between nodes in different clusters. It also rewards links present between nodes in the same cluster and missing links between nodes in different clusters. Here, , represent the individual weights of the four components. Thus, a lower score is better as it means that the internal links and external non-links have more weightage in the model. In a strong model, members within clusters are strongly linked and members in separate clusters are weakly linked; this means members of a cluster are more closely related to each other than they would be to members of another cluster.

After analysing the number of clusters, we created dataframes containing user profiles, along with the questions and topic tags that were discussed within each cluster. We then conducted a content analysis of the characteristics of each cluster, which consisted of conducting qualitative open coding to find themes in users' profiles and topic tags that were present in the various clusters.

Finally, we created a *sociogram* of CSEd SE, that is, a representative diagram of network structure that depicts individuals as *nodes* (i.e., points) and the relationships between them as *edges* or lines (Scott, 2017). To visualize the network clusters, we used a color palette following the spinglass clustering algorithm. To generate the layout of the network nodes, we used the *Fruchterman-Reingold layout algorithm*, which is appropriate for large (but still with less than 1,000 nodes), potentially disconnected networks (Csárdi, Nepusz, & Airoldi, 2016).

Results

RQ1. What are the characteristics of users who contributed to the CSEd SE's core network of interactions?

During the time period of our data collection, contributors to the CSEd SE amassed a total of 559 questions asked by 210 users. There were a total of 2,675 responses to these questions, with 675 different users contributing to the answers. Many users contributed only once and, as mentioned earlier, we excluded them from our *core network of interactions*. We found 257 distinct contributors in the core network: 164 users who asked various questions and 175 who responded to these questions. We calculated descriptive statistics of these core users (see Table 1 below). Note that *degree* here refers to *degree centrality*, a measure of the number of connections that each user has to other users. Thus, degree is helpful in identifying highly connected individuals who are likely to have a lot of influence or hold information in a network. Further, *in-degree* refers to the connections coming towards the node—in this case, it is the number of responses a user might have received on the forum. *Out-degree* refers to the number of responses a user might have made to other users, the connections go outward from a node or user in such a case.

	Degree	In-degree	Out-degree
Mean	9.96	4.98	4.98
Standard Deviation	17.88	9.38	10.35
Median	5	3	3
Minimum	1	0	0
Maximum	146	80	101

Table 1. Characteristics of users in CSEd SE's core network

RQ2. What are the characteristics of the top-10 most connected contributors to the CSEd SE?

The connectedness, or *degree* centrality scores, of the top 10 users ranged from 146 to 37. The mean was 84 and the median was 75. In addition, the number of questions these users asked ranged from 61 to 6. The number answered varied from 292 to 34. These top 10 most connected contributors in the CSEd SE forum had some similarities as well as some differences. In terms of membership duration, these users were mostly similar. There were more variations in the users' location, reputation points, and profession. We also observed that for the most part, this set of users had *reciprocal* (explained below) interactions with each other.

First, the top 10 users belonged to primarily three *professions*: educators in K-12 or college, software professionals, and students. Of the 10, five were educators, three were software professionals, and two were students.

The *locations* of the top users varied somewhat. The majority of the users, six, were from the United States. One was from the United Kingdom. One other was from Israel. The other two had not specified the locations they were from.

The *reputation scores* of the users ranged from 16,875 to 1,333. There was a strong correlation of $r = .90$ between the users' degree centrality score and their reputation points. In addition, all three moderators of the forum were a part of the top 10 most connected users.

There was very little variation in the *membership duration* amongst these users. Almost all of the top 10 users, nine, had been members since around the CSEd SE forum started. Thus, at the time of data collection, these users' membership duration varied from one year and four months to one year and two months.

Almost all of the top 10 users had a *reciprocal relationship* (i.e., *mutual engagement*) with other users. This means that a user "A" answered a question asked by another user "B" and "B" also answered a question "A" had asked. Amongst the top 10 users, only one user did not have reciprocal relationships with anyone. This was because they had not asked any questions and only answered others' questions. Of the remaining users, three had a reciprocal relationship with eight others, four had it with seven others and the remaining two had it with five others.

RQ3. Does network clustering occur in the context of the CSEd SE's core network of interactions? If so, what are some characteristics of these clusters?

We found that there indeed was statistically significant evidence of clustering in the context of the CSEd SE's core network of interactions, meaning that some users in the network are more closely related to each other than they are with others. There were a total of eight clusters in a total of varying sizes. Some of the characteristics of the clusters are given in Table 2 below.

Cluster number	Cluster size or number of members in each cluster	Number of top 10 users present	Top three topic tags (Percentage of presence)
1	43	1	Resource Request (8%), Student Motivation (7%), Lesson Ideas (5%)
2	42	2	Lesson Ideas (7%), IDE (7%), Labs (5%)
3	4	0	N/A
4	2	0	N/A
5	46	1	Best Practice (18%), Curriculum Design (9%), Student Motivation (8%)
6	53	3	Curriculum Design (7%), High School (7%), IDE (7%)
7	42	2	Curriculum Design (7%), Web Development (5%), Lesson Ideas (4%)
8	23	1	Undergraduate (14%), Best Practice (9%), Curriculum Design(7%)

Table 2. Characteristics of various clusters of users in CSEd SE

In addition, we created a visualization of the network, with different colors representing different clusters in the CSEd SE forum (Figure 1). We observed that users in the periphery only have a few connections. The users in the center of the cluster are more connected and have a higher *degree* centrality than those in the edges of the graph. We also observed that although users may belong to certain clusters, they still have multiple interactions with members from other clusters. This means that members belonging to different clusters are not isolated from each other. Indeed, of the top 10 users, each user had connections with members from almost every other cluster.

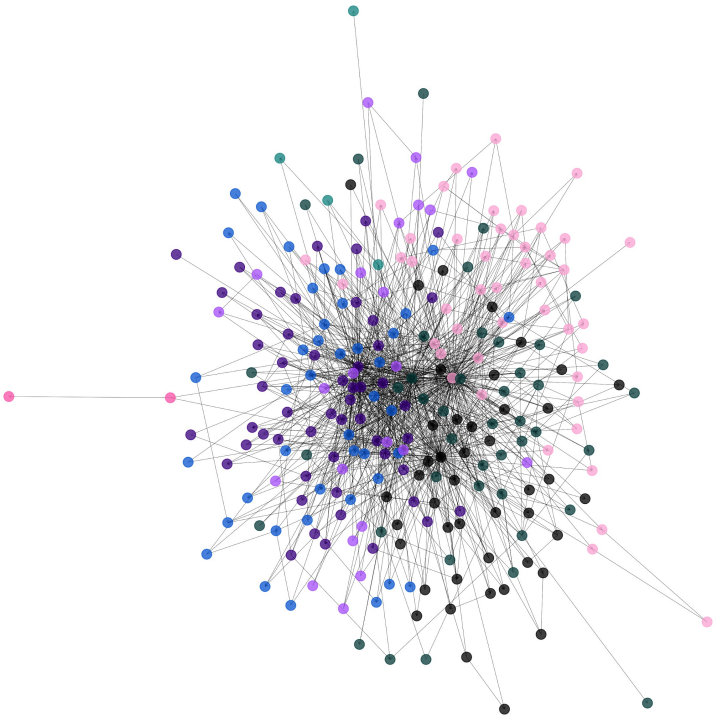

Figure 1. Visualization of various clusters of users in CSEd SE

Discussion

The answers to our first research question revealed that there was a wide range of user characteristics of the members who participated in the "core interactions" in CSEd SE. The highest degree centrality score for a user was 142; the lowest was one. The median (five) and mean (10) were low as well. This means that in the forum, there were only a few users who contributed to the most number of interactions of asking and answering questions. This trend of *participation inequality* is seen in many online discussion forums and social media sites (Selwyn, 2000). A comparison of various forums could reveal which have higher levels of interactions, with more users contributing to conversations as opposed to a few users contributing to most conversations. Teacher educators might find such information useful when directing teachers to various forums as a part of their professional development.

The answers to our second research question revealed that there were both similarities and differences between the top-10 most connected users. It appeared that longer membership duration was an important similarity that the most active users had. Although this is not proof of a causal relationship, it fits well with prior research findings that longtime members in a forum tend to be more active (Graham & Wright, 2014; Sonnenbichler, 2010). In addition, reputation points of users were highly correlated with their degree centrality scores. This too seems to be in line with prior findings that users with high reputation points tend to also be active in terms of both asking and answering questions (Movshovitz-Attias et al, 2013). There has not been much work done regarding the location and professions of users affecting their level of activity in an online forum. Variation in both these factors means that the forum is mostly democratic and there is not a monopolization from one country or profession. Finally, there was a clear sense of mutual engagement amongst the top 10 users. As mentioned earlier, mutual engagement is an important aspect of a community of practice; it is an important contributor to building a community (Wenger, 1998a). Evidence that mutual engagement is possible in an informal online forum means teachers who are singletons in their schools may have opportunities to form communities outside of their classrooms and schools if they have access to good online educator forums.

Finally, our third research question brought to light that there was evidence of clustering in CSEd SE. The different clusters varied in sizes, and, in addition, the top 10 users were spread out amongst these clusters. The clusters which did not have any of the top 10 users were the smallest in size. Sonnenbichler's report (2010)

highlighted the fact that certain users in an online forum act as *leaders* with a high degree centrality score, active participation, large personal network, and the capacity to be a trendsetter within the forum. Indeed, according to Sonnenbichler, *leaders* often have roles of moderators in online forums. Perhaps due to these reasons, the clusters with no *leaders* or active users tended to be the smallest in size. This information might be useful to practitioners, as it is indicative that having some leaders or moderators might be useful for online forums.

Apart from cluster size, the topics each cluster focused on also varied slightly. These popular cluster topics seemed to relate to the topics that the top 10 users participated in, but only to a small extent. For instance, the top contributor of cluster eight had "Grading," "Undergraduate," and "Curriculum Design" as their top topics. But grading scored quite low in popularity (rank 6) in cluster eight's top topic tags. This possibly means that although each cluster may have top contributors and leaders, these users may not solely drive the topics the cluster focuses on the most. Once again, evidence alludes to a mostly democratic forum dynamics.

Thus, overall, CSEd SE has many elements of a community of practice. For instance, there are clear boundaries in terms of membership to the forum. The forum has a unique purpose: to build knowledge and dialogue in the domain of CS education. Further, the discussions on the forum do not seem to be a unidirectional venture—there is evidence of mutual engagement between users. Finally, the presence of moderators and editors for maintenance of the forum, the scope of democratic elections for the selection of moderators, and reputation points based on user-generated voting system, gives CSEd SE the traits of a community of practice.

Limitations and Delimitations

Due to time constraints and the scope of this paper, we did not communicate directly with CSEd SE users; however, future research should consider surveying or interviewing contributors to obtain a more nuanced understanding of their activity on and experience with the forum. For instance, with our current methods, we were unable to determine who *upvotes* which post in CSEd SE—this information can be obtained only by asking users. Further, in order to more deeply study mutual engagement, future researchers should consider including data from the comments section in addition to question and answer data.

Conclusion and Future Research

We conclude from this study that online forums, such as the CSEd SE, may be promising places for educators to form communities and discuss various topics relating to teaching and learning because we found the forum to be democratic in nature in terms of professions and locations of the members. In addition, the forum exhibited some characteristics of a community of practice such as having a joint enterprise of advancing discussions in various topics relating to CS education and the presence of mutual engagement. That being said, it is also important for researchers and practitioners to note that not all online forums may be suitable for educators. Certain features of online forums, such as the capacity to have moderators, official membership, having the ability to vote and rate users with reputation points, the presence of highly reputed "leaders," and features that enhance mutual engagement might affect the extent of participation in the forums. More research is needed to investigate the characteristics of forums that may be most suitable for educators to have to enhance their professional networks.

Finally, various practitioners, particularly teacher educators, could employ the methodologies used in this paper to identify top users, presence of clusters, and top cluster topics in the context of the forums or teacher communities they might be in charge of. Early identification of active reputable users and sub-groups or clusters focusing on certain topics might be useful for online teacher development initiatives.

References

Brown, N. C. C., & Kölling, M. (2013, August). A tale of three sites: Resource and knowledge sharing amongst computer science educators. In *Proceedings of the ninth annual international ACM conference on International computing education research* (pp. 27-34). ACM.

Computer Science Educators Stack Exchange (n.d.) Retrieved January 3, 2019 from https://cseducators.stackexchange.com/

Csárdi, G. (2018). igraph: Network analysis and visualization (Version 1.2.2) [R package]. Retrieved from https://CRAN.R-

project.org/package=igraph

Csárdi, G., Nepusz, T., & Airoldi, E. M. (2016). *Statistical network analysis with igraph*. Springer. Retrieved from https://sites.fas.harvard.edu/~airoldi/pub/books/BookDraft-CsardiNepuszAiroldi2016.pdf

Graham, T., & Wright, S. (2014). Discursive equality and everyday talk online: The impact of "superparticipants". *Journal of Computer-Mediated Communication*, *19*, 625-642.

Hur, J. W., & Brush, T. A. (2009). Teacher participation in online communities: Why do teachers want to participate in self-generated online communities of K–12 teachers?. *Journal of research on technology in education*, *41*, 279-303.

Kadushin, C. (2012). *Understanding social networks: Theories, concepts, and findings*. New York, NY: Oxford University Press.

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., & Faloutsos, C. (2013, August). 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* (pp. 886-893). ACM.

Ni, L., Guzdial, M., Tew, A. E., Morrison, B., & Galanos, R. (2011, March). Building a community to support HS CS teachers: the disciplinary commons for computing educators. In *Proceedings of the 42nd ACM technical symposium on Computer science education* (pp. 553-558). ACM.

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.0) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E, 74* (1), 016110.

Robinson, D. (2015). stackr: An R package for connecting to the Stack Exchange API [R package]. Retrieved from https://github.com/dgrtwo/stackr

Ryoo, J., Goode, J., & Margolis, J. (2015). It takes a village: supporting inquiry-and equity-oriented computer science pedagogy through a professional learning community. *Computer Science Education*, *25(4)*, 351-370.

Scott, J. (2017). *Social network analysis*. London, UK: SAGE.

Selwyn, N. (2000). Creating a "connected" community? Teachers' use of an electronic discussion group. *Teachers College Record, 102*, 750-778.

Sonnenbichler, A. C. (2010). A community membership life cycle model [technical report]. Karlsruhe Institute of Technology. Retrieved from https://arxiv.org/pdf/1006.4271.pdf

Stack Exchange. (n.d.). About - Stack Exchange. Retrieved January 3, 2019 from https://stackexchange.com/about

Trust, T., Krutka, D. G., & Carpenter, J. P. (2016). "Together we are better": Professional learning networks for teachers. *Computers & Education*, *102*, 15-34.

Wenger, E. (1998a). *Communities of practice: Learning, meaning, and identity.* New York, NY: Cambridge University Press.

Wenger, E. (1998b). Communities of practice: Learning as a social system. *Systems thinker*, *9*(5), 2-3.

Wenger, E., McDermott, R. A., & Snyder, W. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Cambridge, MA: Harvard Business Press.

Yadav, A., Gretter, S., Hambrusch, S., & Sands, P. (2016). Expanding computer science education in schools: understanding teacher experiences and challenges. *Computer Science Education*, *26*(4), 235-254.