

Analyzing User Participation Across Different Answering Ranges in an
Online Learning Community

by

Ritesh Reddy Samala

A Thesis Presented in Partial Fulfillment of
the Requirements for the Degree
Master of Science

Approved November 2015 by the
Graduate Supervisory Committee:

Erin Walker, Chair
Kurt Vanlehn
Gary Hsieh
Jon Wetzel

ARIZONA STATE UNIVERSITY

December 2015

ABSTRACT

Online learning communities have changed the way users learn due to the technological affordances web 2.0 has offered. This shift has produced different kinds of learning communities like massive open online courses (MOOCs), learning management systems (LMS) and question and answer based learning communities. Question and answer based communities are an important part of social information seeking. Thousands of users participate in question and answer based communities on the web like Stack Overflow, Yahoo Answers and Wiki Answers. Research in user participation in different online communities identifies a universal phenomenon that a few users are responsible for answering a high percentage of questions and thus promoting the sustenance of a learning community. This principle implies two major categories of user participation, people who ask questions and those who answer questions. In this research, I try to look beyond this traditional view, identify multiple subtler user participation categories. Identification of multiple categories of users helps to provide specific support by treating each of these groups of users separately, in order to maintain the sustenance of the community.

In this thesis, participation behavior of users in an open and learning based question and answer community called OpenStudy has been analyzed. Initially, users were grouped into different categories based on the number of questions they have answered like non participators, sample participators, low, medium and high participators. In further steps, users were compared across several features which reflect temporal, content and question/thread specific dimensions of user participation including those suggestive of learning in OpenStudy.

The goal of this thesis is to analyze user participation in three steps:

- a. Inter group participation analysis: compare pre assumed user groups across the participation features extracted from OpenStudy data.
- b. Intra group participation analysis: Identify sub groups in each category and examine how participation differs within each group with help of unsupervised learning techniques.
- c. With these grouping insights, suggest what interventions might support the categories of users for the benefit of users and community.

This thesis presents new insights into participation because of the broad range of features extracted and their significance in understanding the behavior of users in this learning community.

Dedicated to my dear grandma,
Susheela Mekala

ACKNOWLEDGMENTS

I am very much thankful to Dr. Erin Walker for being a great advisor. The long association has been an important learning period for me.

I am grateful to my committee members Dr. Kurt Vanlehn, Dr. Gary Hsieh and Dr. Jon Wetzel for their acceptance to be a part of the committee and providing their valuable feedback.

I sincerely thank Dr. Preetha Ram, CEO of OpenStudy, for providing OpenStudy's data to perform research and making the process much easier.

I would like to take this opportunity to thank the whole dragoon team at ASU, especially project manager Dr. Jon for his support and arranging flexible schedule to balance my thesis and assistantship. I am also thankful to my friends Nichola Lubold, Pradeep Chaudhari and Sachin Grover for their technical and research support at various stages of the work.

I am very grateful to my parents and family for their encouragement and support through my thesis.

TABLE OF CONTENTS

	Page
LIST OF TABLES.....	viii
LIST OF FIGURES.....	ix
CHAPTER	
1. INTRODUCTION.....	1
1.1 Introduction to the Problem.....	1
1.2 Motivation	4
1.3 User Answering Ranges	5
1.4 Research Questions.....	9
1.5 Outline of the Document.....	10
2. BACKGROUND STUDIES AND RELATED WORK.....	11
2.1 Social Question and Answer Communities (Social Q and A)	11
2.1.1 Background of Social Q and A's	11
2.1.2 Research in Social Q and A's	13
2.1.3 Learning in OpenStudy.....	13
2.2 A Review on User Participation	18
2.2.1 Defining User Participation	18
2.2.2 Analyzing User Participation.....	20
2.2.3 Importance of Participation (Yahoo answers vs Google Answers) ...	20
2.2.4 User Engagement Groups in MOOCs.....	22
2.3. Educational Data Mining (EDM)	25
2.3.1 Introduction to the Field of EDM	25

CHAPTER	Page
2.3.2 A Review of Unsupervised Learning and Clustering	25
3. DATA AND METHODS.....	29
3.1 Dataset	29
3.2 Data preparation.....	32
3.3 Data Attributes Description.....	34
3.4 Methods.....	36
3.4.1 Clustering.....	36
3.4.2 K-means: an Overview.....	37
3.4.3 Cluster Evaluation Metrics: Silhouette Index.....	39
4. RESULTS AND INTERPRETATION.....	41
4.1 Inter Group Participation.....	41
4.2 Intra Group Participation.....	47
4.2.1 High Participators.....	48
4.2.2 Medium Participators.....	52
4.2.3 Low Participators.....	55
4.2.4 Sample Participators.....	59
4.2.5 Non Participators.....	62
4.2.6 Patterns and Interventions.....	65
4.2.7 Known Limitations.....	67
4.2.8 Future work and Conclusions.....	69
REFERENCES.....	72

APPENDIX	Page
A POST HOC TESTING RESULTS.....	75
A.1 Post Hoc Testing for Inter Group Participation.....	76
A.2 Post Hoc Testing for Intra Group Participation.....	77

LIST OF TABLES

Table	Page
1 F and P Values for Each Feature Across User Answering Ranges from Analysis of Variance.....	37
2 F and P Values for Each Feature Across High Participators from Analysis of Variance.....	44
3 F and P Values for Each Feature Across Medium Participators from Analysis of Variance.....	48
4 F and P Values for Each Feature Across Low Participators from Analysis of Variance.....	51
5 F and P Values for Each Feature Across Sample Participators from Analysis of Variance.....	56
6 F and P Values for Each Feature Across Non Participators from Analysis of Variance.....	58
7 Tukey HSD Test Results for All Answering Ranges.....	66
8 Tukey HSD Test Results for Subgroups in High Participators.....	67
9 Tukey HSD Test Results for Subgroups in Medium Participators.....	67
10 Tukey HSD Test Results for Subgroups in Low Participators.....	68
11 Tukey HSD Test Results for Subgroups in Sample Participators.....	68
12 Tukey HSD Test Results for Subgroups in Non Participators.....	69

LIST OF FIGURES

Figure		Page
1	Yahoo! Answers Points and Levels.....	13
2	Stack Overflow Reputation System.....	14
3	A Look at OpenStudy Mathematics P.....	24
4	Sample Document inside a Mongo DB Collection.....	27
5	Distribution of Users and Questions Answered Count.....	07
6	Cluster Assignments Based on Number of Questions Answered.....	08
7	Histogram of Users in Different Answering Ranges.....	09

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO THE PROBLEM

Learning has taken a whole new dimension since web 2.0. The pedagogical affordances the new web offered are content creation and modification through improved user interfaces, connectivity through social network applications, collaborative environments in the form of forums, discussion groups which imbibe notification systems using advanced web technologies and many more. This has allowed many new applications to see light in the learning domain [1] like

- MOOC's (massive open online courses) which gave free access to variety of courses taught by academicians from elite institutions in the world like the MIT OpenCourseWare, Coursera and EDX.
- Learning management systems like Moodle which enabled new ways to manage courses by giving access to course content like quizzes, lectures as well as evaluating tests online to students irrespective of university or place.

This capability to create, share and consume content online has also paved way for another interesting pedagogy in learning: question and answer websites (Q and A communities). Question and Answer communities are those online places which facilitate users to ask questions and allow others to answer questions. The way questions are answered is either by providing information or through discussions. The different type of question and answer communities are community/social Q and A, expert services, library reference services. Community or social Q and A websites are the most popular type of

communities in this category [2]. They represent those online communities where users participate and contribute on a daily basis in wide variety of topics across different websites, blogs and forums. Some characteristic features of community Q and A websites are rich user interface, user ranking or incentives like medals, categorized questions, user profiles and communication facilitative features like tagging and notifications. Community Q and A websites form an important part of the internet world as evident from their popularity. Alexa.com, a service which ranks websites based on internet traffic has ranked some of the top question answer communities like Yahoo Answers, Stack Overflow both at rank 55, Stack Exchange at rank 165, Quora at rank 158 (<http://www.alexacom>).

Question and Answer communities have been growing as repertoires of knowledge owing to large amount of question and answer content being generated in today's web. They are centers of knowledge and information sharing. Yahoo answers alone has 300 million questions as of 2012 with 7000 questions per hour and 21000 answers on an average [3]. They also form important part of search results in search engines. We can see search queries showing results on first pages from prominent Q and A sites like Yahoo Answers, Wiki Answers, and Quora. Such humongous data in addition to being a valuable resource to learn by catering needs of users on daily basis, also put forwards many challenges like question and answer quality, user contribution and commitment, spam data, irrelevant questions, depth of questions, repetitive questions, community lurkers and user expertise. These have been studied in literature extensively in various kinds of Q and A platforms and generally as well.

Understanding and quantitatively analyzing participation behaviors in online learning communities is an interesting topic to research because gaining deeper insights into participation patterns of users can help to introduce interventions in the communities which can help the community grow in various aspects like improved learning, longer life and serving users better. Question and Answer communities in web exist in different forms, different categories and different scales. For example, a Q and A website can be as huge as yahoo answers with hundreds of categories and subcategories or as simple as a mathematical discussion forum focused on calculus. This might indicate that as the nature and environment of each site varies possibly there are much more complicated types of users and mere count of question asking and answering might not be sufficient to understand how a user is participating in a particular community.

In this research, I present a quantitative analysis of user participation behavior in an open social learning Q and A community called OpenStudy. OpenStudy's learning model is described as open social learning, which blends in open content and an interactive learning community that asks questions and seeks answers [4]. It is a synchronous learning community given the availability of users who are ready to answer questions all the time, instant notification system which notifies users of the replies to their questions.

OpenStudy is unique in terms of environment and the learning settings it offers. Some characteristic features of OpenStudy are

- OpenStudy's primary motto is that users have to provide help but not answers.
- Code of conduct: enforces norms like users should provide help than answers, things which promote a positive environment [5].

- User base who visit the site are mostly high school level students.
- Smart Score System: measures the user performance by compiling skills of problem solving, engagement and teamwork.
- Titles: recognizes users with unique titles to motivate them.
- Instant notification system
- Social activities like fans, messaging

I examine OpenStudy in a quantitative analysis of the temporal, content specific and question/thread specific features. These features provide a multi dimensional insight into user participation. Analyzing these features from all the users and comparing them helps to identify patterns of user participation in OpenStudy. Common patterns among a specific group of users can be used to support the group and thus overall community.

1.2 MOTIVATION FOR THE PROBLEM

Sustenance of the community is the central agenda for any online learning community. User participation analysis, adaptation to the feedback by making necessary interventions is healthy for any community. While participation can be looked up in various dimensions, current research to the best of our knowledge is not broad in question and answer communities. Mere categorization of question askers and question answerers cannot help us provide support or motivate users to perform better. We try to define participation in broader terms in this research and apply unsupervised learning technique k means to find hidden patterns in user participation behavior. These patterns of users which give better insights into nuances of participation are the primary motivation for this research. On top of that we try to understand learning in OpenStudy from the participation

attributes we extracted. We are interested to study new user patterns who are actually trying to learn from the community. Though we do not define specific metrics for learning or analyze content to see if knowledge has been constructed in those question-answer threads, we try to analyze learning from participation point of view. This helps the community to identify learning oriented subgroups and can help them learn better which we discuss more in results section. This is the secondary motivation of the thesis.

1.3 USER ANSWERING RANGES

For the purpose of this research 5 answering ranges of users are defined. Before Listing out these ranges, in this section, importance of answering questions in participation, necessity to divide users in answering ranges and reasons for dividing users into “particular” ranges (listed in a table) are discussed.

Analyzing user participation involves looking into all the actions user performs in the community. Some important user activities/actions in question and answer communities are ask questions, answer questions, comment, reply, favourite, become fans, follow up, tagging, taking responsibilities etc. All the actions are follow up actions in the process of either asking or answering questions. Asking questions is motivated by user needs like learning, getting homework done or preparing for an exam. Question asking, though is an important participation activity, it is effortless as far as users are concerned and has no direct influence on community sustenance whereas getting these questions answered has direct influence on community sustenance. Not all users are motivated to answer questions in the community. Question answering process requires time and effort of a user. Quality answering and healthy community are key factors in a community’s long

life. So, in this research we focus primarily on question answering behavior to analyze participation.

In this research we divide users into groups based on their answering range. Division helps for an in depth analysis compared to an overall study. Overall participation metrics like total number of questions asked, answered, total number of users and average time spent on each question gives an account of group nature and dynamics. Dividing users (in this case by answering ranges) and labelling them helps us to compare, contrast and to get a notion of how participation varies in each of these groups. For example, in OpenStudy, Non participators and Low participators were shown to be better conversationalists than higher answering groups using participation attributes and human observation of thread samples. These kind of insights are not very obvious from undivided and overall analysis of users. Also this kind of analysis helps to cater the needs of each group based on their behavior. In the next section, we explain the division into particular ranges starting with Non participators.

First group under study is the Non participators, who have not answered a single question. Though, by definition Non participators are not answerers and full range of participation features for these users are non existent, these users can be looked in contrast with users in other answering ranges with the help of a small fraction of attributes from the data like questions asked, average replies got to their questions and total duration in the community. The next immediate task was to separate the other users who have answered at least one question. Plots of users and questions answered showed a very large gaps in user answering count as the count increased.

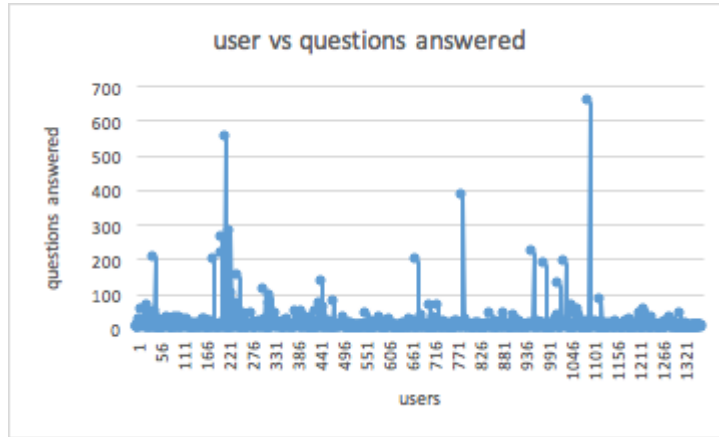


Figure 5: Distribution of Users and Questions Answered Count

When clustering was used to this list of users, users in the last cluster which had high answerers had very low quality with shifting outliers. Increase in the number of clusters resulted in subclusters of high answerers which made the situation more complicated. So, separating high answerers to into a manually defined group has made the process easy and clusters with high quality. This manual separation of high answerers included several human observations of parameters , especially shift in duration from 40-50 band answerers and 50-60 band answerers was 25 weeks, which is very high. So, above 50 users were split as high participators. This split, interestingly proved healthy to cluster other groups and obtain high quality clusters. Following, 1-50 answering band was clustered using k means. Computing several iterations, we decided on clustering combination of (1112, 188 and 46 users) into 3 clusters with average silhouette width 0.78, which indicated a very good structure for these clusters. The other combination was (1172, 137, 37). Both of these combinations repeated in the several times we performed the experiment in. The first combination, which repeated itself several times was also fit into answering ranges of 1-5, 6-20 and 21-50. Thus, clustering was used to obtain these 3 ranges

in addition to non participators and high answerers discussed. Following figure shows the cluster plot.

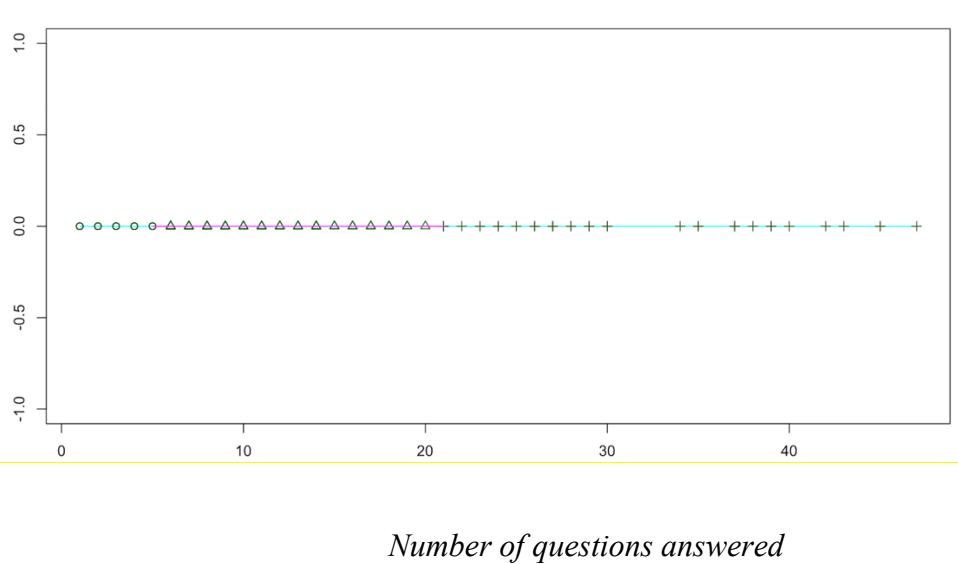


Figure 6: Cluster Assignments Based on Number of Questions Answered

Summary of the ranges is as follows, we use them to describe research questions in the next section.

1. *Non participators(NP)*: User who have not answered a single question.
2. *Sample participators(SP)*: Users who answered less than or equal to 5 questions in the community.
3. *Low participators(LP)*: Users who answered more than 5 questions and less than or equal to 20 questions.
4. *Medium participators(MP)*: Users who answered more than 20 questions and less than or equal to 50 questions.
5. *High participators(HP)*: Users who answered more than 50 questions in the community.

Following graph shows the number of users in each of the answering ranges

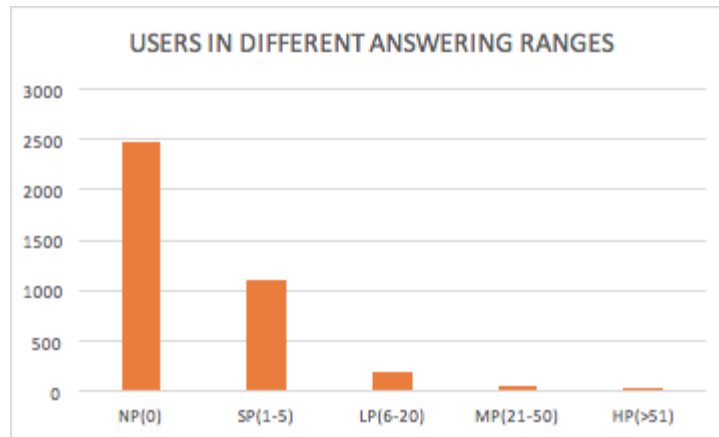


Figure 7: Histogram of Users in Different Answering Ranges

1.4 RESEARCH QUESTIONS

There are 3 research questions we would like to address in this work

1. What are the inter group differences between different category of users like non-participators, samplers, low participators, medium and high participators.
2. What are the intra group differences? What are the different kinds of user sub groups we can find in each major group and how do they vary among the participation attributes we describe in the document?
3. What could be possible interventions we can make after analyzing each subgroup and possible design changes respectively so that user experience can be improved in terms of learning, contribution, and ultimately the sustenance of the community?

1.5 OUTLINE OF THE DOCUMENT

The rest of the document starts with some background work on social Question and Answer communities which is most relevant category to the learning community we are experimenting on, then learning in online communities is discussed followed by detailed description of user participation problem, defining participation, importance of the participation and our take on participation. We conclude the background work with a review of educational data mining literature and the kind of problems the field addresses in online learning communities, then give a detailed description on k means clustering and silhouette cluster evaluation techniques which are the data mining concepts used for this study.

The next half of document discusses OpenStudy data, participation attributes of OpenStudy, data extraction and preprocessing of OpenStudy data to relational database style structure followed by the exact method we used to extract participation subgroups in OpenStudy, then we try to present an overview of the results got from our study.

We conclude the document with references for future work, interventions possible for steady community sustenance.

Chapter 2

RELATED WORK AND BACKGROUND STUDIES

In this section we briefly cover ground on social Q and A's from their origin, current research, learning in Social Q and A (specifically OpenStudy). We then move onto provide an overview of user participation, briefly discuss 2 background papers which inspired our work and finally conclude with an introduction to Educational Data Mining (EDM) and do a survey on usage of unsupervised learning technique in EDM literature.

2.1 SOCIAL Q and A's

Social question and answer communities, sometimes called community Q and A's are a Web 2.0 enabled systems where users create content collaboratively while they ask questions and answer other's questions [6]. Users exploit social Q and A's to feed their information needs across various topics/categories and in the process rate, comment and share the content.

2.1.1 BACKGROUND OF SOCIAL Q and A's

Gazan has provided a detailed description on the evolution of social Q and A's [6]. History of social Q and A's dates back to purposes of basic web information processing systems. Search Engines were used to crawl and present relevant information based on user input in form of natural language (mostly English) key words. We can observe more or less similar phenomena in social Q and A's in the current web era, retrieving similar questions based on user input keywords is an important part of these communities. These information retrieval systems were followed by some specific research on QA focused answer retrieval systems from the user queries. Then comes the precursors of today's social Q and A's that are the Usenet newsgroups.

Usenet system is a collection of discussions/messages from users across different world locations. They are similar to public bulletin board organized hierarchically. Usenet groups comprised of threaded user conversation in various topics. They were a popular internet service before the world wide web and were non commercial in nature. Usenet system was developed by Tom Truscott and Jim Ellis from the general purpose UUCP dial-up architecture. They appear similar to today's discussion forums on the web but principally contrast on the network architecture. Unlike today's world wide web where information services or resources are held on a properly administered central server which we call websites, Usenet systems are non centralized and are distributed among a constantly changing conglomeration of servers that communicate within each other through news feeds. Users in order to post to a news group have to submit their messages to a local server which is connected into this conglomeration by an organization, internet service providers, university, employers etc. These systems were succeeded by an umbrella of applications which more or less resembled and built on Usenet systems which were made by possible by Web 2.0 [7].

With the advent of Web 2.0 there were many sites, systems and applications that reflect the central idea of community question and answering. The idea of web based question answering community implies self evolved systems which gave certain capabilities to the users to post questions, add comments, give replies, sharing the questions, forming groups etc. all through a relatively rich user interfaces. Each community organized the systems in their own styles and imposed their very set of norms. For example, Yahoo Answers is a large community focusing on multiple categories whereas stack

overflow focusses on a more computer programming oriented community with strict rules in terms of question duplication, quality of question asking. Google answers was a paid site which provided remuneration to set of answerers unlike most of communities which are free and open to public. The first Social Q and A was launched by a South Korean company and was called Knowledge-iN while Answerbag was the first in united states. The notion of social Q and A became popular with the advent of Yahoo Answers which is the most widely used social Q and A. Other popular sites are Stack Exchange, Quora, Wiki Answers etc.

2.1.2 RESEARCH IN SOCIAL Q AND A's

Gazan identifies three primary areas of research in social Q and A, they are user motivation and behavior, information quality assessment, and design and technological factors impacting participation. Research review further continues in this work stressing on details of frame works to understand value of information, structuration theory-communities of practice, meta discussions in social Q and A sites, central motivations for user participation on social information spaces, intrinsic and extrinsic motivations, user satisfaction, collaboration, user reputation. This work is a complete review of work in social Q and A and also proposes future research directions.

2.1.3 LEARNING IN SOCIAL Q and A's (OPENSTUDY)

Understanding and quantitatively analyzing learning specific behaviors in online learning communities is an interesting topic. By definition “**Learning** is the act of acquiring new, or modifying and reinforcing, existing knowledge, behaviors, skills, values, or preferences and may involve synthesizing different types of information” [27]. Q and A

communities in web exist in different forms, different categories and different scales. For example, a Q and A website can be as huge as Yahoo Answers with hundreds of categories and subcategories or as simple as a mathematical discussion forum focused on calculus. So, in some communities learning is mere gaining of facts, sometimes it is discussion on certain opinions, events, sometimes it is more technical like in the case of Stack Overflow. So, learning is very subjective in various levels in Question and Answer websites.

In this research we analyze Q and A participation behavior which are related to processes beneficial for learning. We present a quantitative analysis of user participation behavior in an open social learning Q and A called OpenStudy. We do so by extracting various attributes of user participation which might be suggestive of learning. User participation behaviors like answering questions in detail, prompting high discussion while answering, trying to help more number of users, being consistent in helping users, not giving mere answers to homework questions, asking questions which prompt higher discussion, answering questions similar to asked questions and helping users in related topics could be more fruitful than just giving answers, answering questions by addressing more topics are all supporting behaviors of learning. Other way to look around is to keep a check on lurking behaviors, one-time site users, user who are inactive for major periods of duration, mere consumption, users who answer lot of questions in expert level with one or very few replies All these behaviors contradict the learning goal. Though these behaviors are quite natural in a community setting, identifying such behaviors in advance could help the site designers to propose interventions for recommendations or redesigns which we discuss more in conclusions and intervention section of this document.

In this research we try to identify such behaviors with the help of participation attributes extracted in an open social learning Q and A called OpenStudy. we do this by considering following features

Temporal features:

- a. Total duration on the community,
- b. Number of active weeks (time units) in the first year of answering,

Learning can be influenced by the user activity time and his duration on the community. People who stayed for longer duration are more likely to contribute more, though there are people who stay for longer duration but very less of contribution. The total duration and active weeks attributes help us to capture this information.

- c. Best participation point: the time unit where user has answered maximum number of questions on a scale of 0-1,
- d. percentage of questions answered (we call this best given).
- e. top week of participation and number of questions answered in that week

Best participation point and best given features give us an idea of behaviors where users are peaking in their journey and what percentage of answers they give while peaking. Top week and performance in top week complement these features. We also collect previous and after performance metrics which reveal the consistency in user participation.

Content specific feature:

- a. Topics addressed on an average in a week
- b. Similarity of topics in questions asked and answered.
- c. Number of words in the questions answered
- d. Number of words in replies given

While it is hard to grasp the semantics and the knowledge transfer in a thread, some quantitative features might help us to grasp this information on whole. For example, the number of topics across which user has been participating on an average in his duration might help. Less number of topics might imply focus on few and limited concepts while broader topic range indicates broader contribution, also similarity of questions asked and questions answered across topics might indicate and provide us a measure that user has not been answering randomly but has been asking questions and answering on relevant topics to an extent in association with other features like the total number of questions asked and answered.

Question and thread specific:

- a. Questions asked: number of total questions a user has asked
- b. Replies got for the questions asked [Length of the thread]: average number of replies user has got to his questions
- c. Questions answered: number of total questions answered
- d. Average replies given to a question: average replies given in answering questions in the community.

Thread specific content like number of replies generated from other users on a question user asked might indicate a higher discussion which is an essential indicator of knowledge transfer or construction within the thread. Also number of replies a user is giving in association with question/answer complexity for which count of words are a potential measure while he answers a question might indicate his contribution towards a question. So, we extract the discussed features.

Summary of features pertaining to participation and learning

Properties (Features)	Participation	Learning
Question asking behaviour (Questions Asked, average replies got)	Questions asked and average replies got reflect the one important side of participation of question asking behavior	Similarity of questions asked and answered might be an indicator of knowledge gained from community and learning behavior. High replies also indicate good conversationalist which depict user interest in learning
Question answering behavior (Questions answered, average replies got, answer words)	Question answering behavior using these attributes is the central participation behavior in this research. We use question answered to divide users into groups to further investigate participation. Good number of replies given in higher answer words indicate quality participation	Quality participation might have learning implications. It reflects users commitment in contributing towards learning and not giving mere answers
Consistent participation (Total Duration, Active weeks, Best participation point/ top performing week, best given/top performance in a week, previous and after participation)	Consistent participation can be identified by examining all these specific features. Identifying consistency helps to detect initial enthusiasm in participation, early and late participators, lazy participation.	High consistency might be an indicative of learning but experts who have high consistency were also found to give mere answers in shorter replies
Topical analysis (Average topics per week, topical similarity of questions asked and answered)	Topical analysis can be used to detect high participation. Better participators answered in more topics in an average	Topical similarity might be used to detect learning. As discussed in results section high similarity users were found to answer questions related to the one's they asked

		generally in high quality conversations.
--	--	--

2.2 A REVIEW ON USER PARTICIPATION

In this chapter, the primary goal is defining user participation and observing the importance of user participation. Background studies are reviewed briefly which have motivated this research. Later user participation in OpenStudy is discussed.

2.2.1 DEFINING USER PARTICIPATION

Defining participation in online learning communities is site specific in certain terms as it depends on the type of community, agenda/norms of the community, design etc. A simple example could be a comparative between Yahoo Answers and Stack Overflow [8][9].

Yahoo Answers employs a system of points and levels. User's actions measure user's points/levels. Scoring and points are the information gathered from official yahoo answers website [8] details this system in the points and levels figure.

Points and Levels	
Action	Points
Begin participating on Yahoo Answers	One Time: 100
Ask a question	-5
Choose a best answer for your question	3
Answer a question	2
Self-deleting an answer	-2
Log in to Yahoo Answers	Once daily: 1
Have your answer selected as the best answer	10
Receive a "thumbs-up" rating on a best answer that you wrote (up to 50 thumbs-up are counted)	1 per "thumbs-up"
Receive a violation	-10
Levels	

Figure 1: Yahoo! Answers points and levels

Source: https://answers.yahoo.com/info/scoring_system

It is obvious from their point and levels system that Yahoo Answers is more focused on users being active answerers. As asking question will reduce your score which shows the site motivates users to answer more. Also it notes that users can participate more as they get to higher levels scoring higher points.

Stack overflow on the other hand encourages users to be ask high quality questions and provide high quality answers. User participation metric in stack overflow called reputation is influenced by various factors as detailed in their official website [9].

You can earn a maximum of 200 reputation per day from any combination of the activities below. [Bounty awards](#), [accepted answers](#), and [association bonuses](#) are not subject to the daily reputation limit.

You gain reputation when:

- question is voted up: +5
- answer is voted up: +10
- answer is marked "accepted": +15 (+2 to acceptor)
- suggested edit is accepted: +2 (up to +1000 total per user)
- bounty awarded to your answer: + full bounty amount
- one of your answers is awarded a bounty automatically: + half of the bounty amount ([see more details about how bounties work](#))
- site association bonus: +100 on each site (awarded a maximum of one time per site)

If you are an experienced Stack Exchange network user with 200 or more reputation on at least one site, you will receive a starting +100 reputation bonus to get you past basic new user restrictions. This will happen automatically on all current Stack Exchange sites where you have an account, and on any other Stack Exchange sites at the time you log in.

You lose reputation when:

- your question is voted down: -2
- your answer is voted down: -2
- you vote down an answer: -1
- you place a bounty on a question: - full bounty amount
- one of your posts receives 6 spam or offensive flags: -100

Figure 2: Stack Overflow reputation system

Source: <http://stackoverflow.com/help/whats-reputation>

While we do not go deep into participation metrics because the central idea is to make the point that each site measures participation differently according to their own

agenda/norms etc. These metrics are a motivation to user to participate on his end and check his progress with the community.

Given the participation metrics, it is important to analyze in the real time how participation is affecting the community. Participation is a complex variable, it can be as simple as count of questions answered in a Q and A community, Number of home works finished in a MOOC or as complex as topic modelling of question/answer content generated, Longitudinal improvement metrics.

2.2.2 ANALYZING USER PARTICIPATION

Irrespective of how we choose to measure participation, analysis of participation helps in understanding hidden factors/patterns of user behavior which are not obvious from a user's score and this requires examining the participation from various dimensions. Researchers have analyzed participation to understand sustenance/failure of a community, to propose interventions and redesign community, to expand the pre existing views of participation in literature, etc.

2.2.3 IMPORTANCE OF USER PARTICIPATION

(YAHOO ANSWERS vs GOOGLE ANSWERS)

User participation model is core to the sustenance of a community and could decide the fate of the community. In their work shah et al [10] compare participation in Q and A's as important as indexing, retrieval and ranking are to a search engine. They compare participation across yahoo and google answer sites and argue that google answers was a failure and has retired eventually because of their policy which controlled user participation. google answers is a paid community which has few answerers who were

experts and while user asked a question, he had to mention the amount he was willing to pay for the best answer. In contrast, highly popular and successful yahoo answers was more participation oriented and open. One striking thing about their analysis was that there was a huge imbalance of contributors and consumers in google answers. The number of people who asked questions were 100 times larger than the preselected 500 experts/contributors whereas the open participation model in yahoo answers showed a well balanced consumers and contributors. This has led to around 63 percent of questions being unanswered in google answers and eventually led to one time consumers and thus dis satisfaction with the service.

Based on Grouping users into different levels of participation based on the metric yahoo answers created, authors have further delved into these groups to present some obvious and on other hand interesting results.

- a. High number of users in lower levels, compared to higher levels
- b. Average of questions answered to asked (contributed to consumed) is very high in higher levels and gradually decreases as the level falls which reflect maturity level yahoo proposes is highly correlated to user contribution.
- c. Correlations between answers and points is too high (0.88) while question-answer / question-points correlation is low (0.22 and 0.30)

The authors analyze that quality of participation in terms of stars a user's question and answers have been given. When questions are considered they propose two metrics, the first one is the average number of stars for their questions and second one is the percentage of questions which received stars. Here the results are quite obvious that

higher level users had more stars on average and got more stars in terms of percentage as well but there was one interesting thing that level 5 users had more percentage of questions starred than level 6.

When answers are considered the metrics were average number of answers and average number of best answers through a single visualization which produced fairly obvious results that users in higher levels have answered more and got more stars to their answers. The other interesting metric in this category is the percentage of answers selected as best answers for users at various levels. This is a quite uniform graph given the number of answers given in lower levels is also low, but it is interesting that even those answers got the stars which was otherwise hidden if just the levels were the criteria.

The takeaway and motivation from this work for our work is mining deep into levels might give a new and different story about participation though quite a lot literature sees the users as contributors and consumers as far as question and answer communities go. Taking from the work we perform analysis among groups of users who were initially grouped by the number of answers they gave. We also expand on the longitudinal aspect of participation which the authors did not include in the current research and propose in future research.

2.2.4 UNDERSTANDING USER ENGAGEMENT GROUPS IN MOOCs

Participation analytics shed light on new categories of users in contrast to established conventions. For example, in Massive Online Open Courses(MOOC's), according to Rene et al. [11] there was a monolithic view of participants that is users who have completed the course and users who discontinued. In their work Rene et al. propose

a classification method to discover new subpopulations and further propose interventions. The notable category of this work is users who merely engage in the community by watching videos but not completing assessment tasks among their trajectory analysis of engagement and disengagement. Using unsupervised learning in such scenarios where users come from different backgrounds and intentions due to the open nature of MOOC's was motivated from similar clustering techniques performed in community colleges to understand user typologies. In order to perform interventions and decision changes along the design of instruction, content, assessment and platform features they compare learner characteristics and behavior. They choose three course in the levels of high school, undergraduate and graduate. The method they choose is giving labels to user participation in certain time intervals (typically weeks), 4 labels which they use are "on track", auditing (viewing the video content but no assessment), "dis engage" (dis engagement), "out" (no participation). user's labels might vary in the weeks to come in the course. For example, a user might be on track in first 2 weeks, then disengage for a while and then only might be auditing, so thus for each user a string of labels is generated, which are then represented numerically for the purpose of applying k means, a most commonly used clustering algorithm. Interestingly the results are similar to the engagement labels. User categories as per their description is Completing, Auditing, Disengaging and Sampling which are descriptive of their behavioral patterns of engagement. Their clustering has produced a good fit as reflected by silhouette width of 0.8 which is very good fit for users in each of the clusters. When they tried to increase k in their clustering, that is the number of clusters, the algorithm further found sub groups in samplers based on their sampling period. The high level clusters gave a good fit. They have also evaluated the clusters from an

educational perspective. They have also experimented with attributes like assignment grades, lecture counts but it seems clusters were not very informative of learning process and pedagogical improvements. Also in the final conclusion they say they would like to examine in terms of more nuanced measures of engagement, user information which we think might have given more in depth understanding of learner trajectories but the good fit could not have been possible. More categories of users might be implying more interventions and tools to support these users. Once they categorized users they try to understand and compare users across categories and courses in terms of survey demographics, geographical location, intention, overall experience, forum activity and streaming index. They present a statistical analysis of each of these features which reflects a deeper dig into user engagement and thus accordingly propose design strategies. For example, let's paraphrase two scenarios. In case of auditing learners they could only be shown videos and avoid the frustrating messages about completing assessments which could disengage them or totally remove assessments. This could happen by predicting auditing learners early from their engagement patterns the clusters reflect. Analyzing the demographics, the dis engaging and sampling users told the reasons for disengagement were their personal commitments, work conflict and course loads. In this case authors suggest a slow pace course or an entirely self paced course work. Similar interesting analysis was made considering cross course and cross cluster comparisons.

While the results are pretty interesting the takeaway from this work is using a data mining technique to dig deeper into user categories based on engagement categories and further analyzing and proposing interventions. We, in our work try to find hidden patterns

inside the clusters and across more broad and subtler attributes of engagement. We could connect to few attributes they used like the number of posts and comments the learners created in the posts.

2.3 A REVIEW ON EDUCATIONAL DATA MINING AND UNSUPERVISED LEARNING

2.3.1 INTRODUCTION TO FIELD OF EDM

Educational data mining is a new emerging field which focuses on application of data mining and machine learning techniques to the data generated from educational settings like online learning resources, learning management systems, universities, intelligent tutoring systems etc. [12]. Increase of computing power and usage of educational technology using computer systems is generating large amounts of data in these settings and the potential of data mining techniques in education is being recently explored and experimented. Baker and Yacef identified four goals of educational data mining

- a. predicting student learning behavior
- b. discovering or improving domain models
- c. studying the effects of educational support
- d. Advancing scientific knowledge about learning and learners

2.3.2 A REVIEW OF UNSUPERVISED LEARNING AND CLUSTERING IN EDUCATIONAL SETTINGS

We believe that in our work, where we use an unsupervised learning technique which is an important part of data mining algorithms to understand the participation behavior of users and suggest possible interventions to improve the experience of users in a learning community is more or less close to the second goal. EDM has broad range of

implications in various kinds of learning settings to address multiple problems, two survey papers from EDM community help us to understand the kind of work possible or being done in the field [13][14].

We focus on clustering, an unsupervised learning technique from data mining/machine learning fields. Unsupervised learning has been used in outside educational applications like collaborative filtering, managing news documents and indexing them according to categories and subcategories, email activities etc. we briefly discuss the kind of work that has been done in various educational settings, why we choose to apply a clustering technique called k means, further explore different type of clustering methods, k means and it's implementation details.

Amershi and Christina propose a user modelling framework using supervised and unsupervised learning which reduces costs of building user models and facilitates transferability across applications. They apply this approach to an intelligent tutoring system which teaches AI algorithms and model student learning during interaction with a learning environment called Adaptive coach for exploration (ACE)[15].

An exploratory learning environment facilitates learning which is student led while potential knowledge discovery happens through generating hypotheses, testing them, finding relevance, making generalizations, reaching conclusions etc. [20]. Pedagogical effectiveness which is influenced by distinguishing student characteristics and learning styles has been mixed in these environments because of the difficulty in finding patterns in unstructured environments. Manually constructing and evaluating student models is difficult, needs collaboration of various people involved and was not proved to be effective. Few researchers attempted to use supervised learning techniques where experts gave labels

based on desired exploratory outcomes. Though this improved the model accuracy it was error prone because humans had to code the labels. Amershi and Conati in their chapter in EDM text discuss various unsupervised learning as a solution in this context to identify common interaction behaviors through the algorithm generated groups and then apply these labels to train a user model [21].

In the participation behavior section, we have explored a MOOC analysis of user engagement pattern from Stanford researchers. Based on this study a similar study was performed not on Coursera but on FutureLearn [22], a learning website which offers MOOC's and according to Ferguson and Clow is underpinned by social constructivist strategy. The interesting part of this work is they performed k means clustering to identify engagement patterns across 4 courses. They found additional seven distinct patterns of user engagement which is more temporal oriented. The groups of users are one among Samplers, strong starters, returners, midway drop outs, nearly there, late and keen completers. We can recall that in the previous study, adding an additional cluster has sub clustered samplers along time of sampling [23].

Researchers have also compared effectiveness of different clustering algorithms for a single data set or a similar educational setting. Shubhendu et al. explore the spectral clustering algorithm [17] in comparison with popularly employed k means algorithm [18] in a year's school data collected from ASSISTments tutor in two schools at Massachusetts and predict student's test scores from features derived from the tutor like number of problems done, correct percentage, correct percent on help questions, time spent, attempts, hints [16]. In another work Beijie et al. compare K means and Latent Class analysis on data from instructional architect, which is a digital library service to identify user groups.

Instructional architect is used by teachers in creating and sharing instructional activities. This work also identifies problems with using clustering in educational settings, they propose to justify the choice of algorithm used, compare with a competing algorithm, discuss about evaluation of the clustering results. This work is a good source of various clustering algorithms used in different problems, in terms of the problems they discussed.

[19]

CHAPTER 3

DATA DESCRIPTION AND METHODS EMPLOYED

3.1 DATASET

Dataset was provided by openstudy.com (<http://www.openstudy.com>). They have given access to chunk of their data for research purposes where we used mongo commands on shell to access the data. In detail data description and processing will be dealt in following steps.

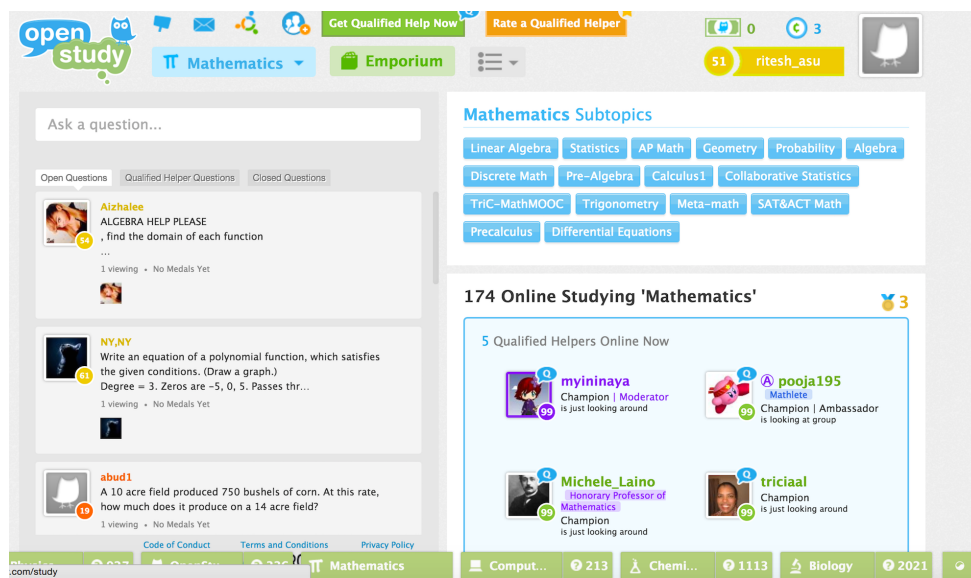


Figure 3: A look of OpenStudy Mathematics page

source: www.openstudy.com

Original data format:

This section gives basic account of original data format and how it was handled initially before starting the actual processing.

Mongo DB:

Data was originally stored in a non traditional database using Mongo DB which is categorized as a No SQL database [24]. Mongo DB is a document oriented database which

uses JSON (JavaScript object notation) based documents which have a dynamic schema and these are called BSON documents. BSON is a binary-encoded serialization of JSON-like documents. It is intended for data storage and network transfer purposes by Mongo DB [25].

JSON vs BSON:

Mongo DB page on JSON and BSON [26] provides the following difference:

JSON is an important open data interchange format widely used on web. JSON supports various data types like numbers, strings, and Boolean values including arrays and hashes.

BSON which stands for binary JSON on the other hand is an extension of JSON used by Mongo DB for binary encoding a JSON file such that it provides additional data types and efficiency for encoding and decoding within different languages.

Understanding the data on Mongo DB:

Mongo Db generally organizes the data into collections which are analogous to tables on a relational database system. collection implies a group of documents in the database. These documents are not enforced any particular schema and also need not have similar fields which reflect the central concept of a No SQL database. This section lists examples of few collections in the database and explains basic commands to execute few common operations.

Collections in OpenStudy:

This section gives an idea on the data we had access to. Collections in OpenStudy are used to store various entities of the community data like

- a. User profile information like first name, last name, date of joining etc.

b. User performance information like medals, smart scores, engagement, team and collaboration metrics.

c. Questions content like body of question, time stamps , status, question askers etc.

d. Reply content like body, timestamps, ownership of replies etc.

e. Groups related data

f. Notifications like medal announcements, reply notifications

g. Other analysis data.

OpenStudy, according to their privacy concerns made sure personal messages and users personal information is nullified and not used for any research purposes.

This is what in a nutshell, collections in OpenStudy had to offer. In later sections of the document we clearly describe how feature set for analysis was formed from the data the collections had.

The following screenshot shows how a sample document inside the notification collection looks like. Information embedded in this particular document include

a. ID of the notification

b. Date

c. From: who produced the notification

d. To whom the notification was delivered

e. Content

f. Status

g. Class of the notification

```

> db.notifications.findOne()
{
  "_id" : ObjectId("4db82b693b057d737d39f1b8"),
  "associatedId" : ObjectId("4d670b595f368b0bd5e3b5b0"),
  "date" : ISODate("2011-02-25T02:03:37.887Z"),
  "from" : [
    "marlondude",
    "marlondude",
    "mathfreak",
    "mathfreak",
    "mathfreak",
    "mathfreak",
    "marlondude",
    "mathfreak",
    "marlondude",
    "mathfreak",
    "marlondude",
    "marlondude",
    "mathfreak",
    "marlondude",
    "mathfreak",
    "marlondude",
    "marlondude",
    "mathfreak"
  ],
  "jsonClass" : "ReplyNotification",
  "read" : true,
  "text" : "what is 20 fl oz= c?\n",
  "to" : "*math*_sparkle"
}

```

Figure 4: sample document inside a Mongo DB collection, source: OpenStudy

3.2 DATA PREPARATION

Data preparation refers to that process of collecting, cleaning, and consolidating data into one file or data table for use in analysis. In our case of preparing data for analysis we break into 3 distinctive steps for better understanding

Step 1: Data transporting

Step 2: Data restructuring

Step 3: clean, consolidate and simplify for analysis

In this section we explain each step with necessary detail to get an idea of how data was transformed into convenient format which is ready to analyse

Step 1:

Data Transporting:

Data transporting implies moving the data from server onto a local machine in the process of simplification for analysis. Transporting involves

Creating data dumps of collections on the server in BSON formats:

As discussed mongo provides capabilities of creating BSON dumps for transportation facilitation. Secure copying the files onto a local machine which also hosts a local Mongo DB server. Recreating the Mongo DB on local machine after re converting them to JSON formats.

Final steps in this process points to writing scripts which query the data and creates new SQL tables, a new format for representing the data for various conveniences. Further restructuring will be briefed in step 2.

Step 2:

Data Restructuring:

Data restructuring simply means creating a new schema for relational database to be created. All necessary data was put into users table Questions table and replies table

This step helped to find basic trends in the data which gave us sense of the nature of community. For example like number of users who asked one question only , number of users who answered more than 50 questions to get a sense of relation between percentage of contributions and user count.

Step 3:

Clean, consolidate and data simplification:

Users who were removed from the community due to various reasons were also deleted from the new SQL database. All the dependent data items like questions, replies were removed to preserve the consistency. We filtered out the whole database to create smaller tables which corresponded to chemistry subgroup of OpenStudy. In terms of simplification, this helped the scripts to run faster and made the data processing handy because the primary database was huge to process in feasible times. We then created a final table with the attributes chosen to perform the analysis which we describe as follows

3.3 DATA ATTRIBUTES DESCRIPTION

1. **Questions asked:** total number of questions each user has asked in the period of his first year

2. **Average Replies got:** average replies user has got on his questions excluding his replies to his own thread. we divided the total replies by total number of questions asked

$$\text{Average replies got} = \text{total replies got} \div \text{total questions answered}$$

3. **Questions answered:** total number of questions user has answered in the chemistry subgroup

4. **Average Replies given:** average number of replies user has given excluding replies to his own questions

$$\text{Average replies given} = \text{total replies given} \div \text{total questions answered}$$

5. **Total duration:** total lifetime of user in the group. we calculated this by sorting out user activity timestamps to get first activity and last activity timestamps

6. **Active weeks:** In the first year of participation, out of those 52 weeks, the count of the weeks user has answered at least one question. Looping in first activity timestamp plus week timestamp has given us the number of active weeks.

7. **Question words:** average total number of words in the questions user has answered

$$\text{Question words} = \text{total question words} \div \text{questions answered}$$

8. **Answer words:** average total number of words in the replies user has given

$$\text{answer words} =$$

$$\text{total words from all replies given to a question} \div \text{questions answered}$$

9. **Best participation point:** point from scale of 0 to 1 when user has given his best in the first year.

$$\text{BPP} = \text{top week/duration}$$

10. **Best given:** percentage of best given in the top week

$$\text{BG} = \text{questions answered in top week/total questions answered}$$

11. **Previous performance:** total number of question answered before the top week

12. **After performance:** total number of questions answered after the top week

13. **Top week:** week where highest questions were answered

14. **Performance in top week:** number of questions answered in top week

Topics per week: average number of topics user has addressed in his answering life per week

15. **Similarity:** cosine similarity between topical vector of questions asked and questions answered.

Topic modelling method and cosine similarity will be discussed along the methods section along with clustering methods.

DATABASE DESCRIPTION:

The database under description, OpenStudy, was not completely used in this research. Only a chunk of this data, “chemistry” subject group was utilized. All the questions, replies, users and other information pertaining to chemistry was extracted for the purpose of this study. Some important chemistry data statistics are

Number of questions asked: 9047

Number of users: 3863

Users who answered at least one question: 1378

Users who asked at least one question: 3102

3.4 METHODS

3.4.1 CLUSTERING

Clustering is the process of grouping similar objects in the data to groups/sub groups. This in turn helps to find the structure otherwise not implicit in the data. Unlike classification where the label is known pre hand and data is trained accordingly, clustering completely starts without any labels and groups data using various types of algorithms.

Clustering has application in various domains like biology, astronomical data, health, education, social network analysis, e commerce, recommendation systems, climate and other time series data.

Various clustering techniques are presented in literature [27]. prominent categories are present based on the approach of clustering the objects.

- a. Connectivity based clustering
- b. Centroid based clustering
- c. Distribution based clustering

d. Density based clustering

For the purpose of this research we employ a widely used algorithm called k means which is a centroid based clustering technique, which is reviewed in detail in the following section.

3.4.2 K-MEANS: AN OVERVIEW

K-means clustering is most popular and frequently used method in cluster analysis. As seen in the background section K means has broad range of application. Kumar et al. in their data mining cluster analysis book present an excellent and simple overview of k means, it's limitations and extensions [27].

We start with initially choosing k centroids while k is a user specified number. selection of k can be done with various methods like using scree plot, sometimes it is subjective, tried and tested based on domain knowledge. Once k is chosen, k random centroids are initialized and each point in the dataset is assigned to the closest centroid. The centroid is updated based on the points assigned to the cluster and this process is repeated until convergence , that is there are no more points changing across clusters and the centroids remain intact. following we describe algorithmic steps described in most standardized source

Algorithmic steps:

1. Select k initial centroids
2. Repeat
3. Form k clusters by assigning each point to its closest centroid.
4. Re compute the centroid of each cluster.
5. until centroids remain intact

Note: 3 and 4 are repetitive steps

While the condition is that points do not change clusters often restrictions are laid practically like n percentage of points change clusters.

An important part of this algorithm is point assignment to the centroid, for which a distance metric has to be chosen. The most commonly used is the Euclidean distance in the Euclidean space (higher dimensional space in our case). Other similarity metrics used are Manhattan distance, cosine similarity, Jaccard similarity. we used the simple Euclidean distance in our case.

Re computing the centroids is another key step in k means algorithm, in re computing generally an objective function is laid like minimizing the sum of squared distances of points to the closest cluster centroids. In case of Euclidean distance sum of squared error (SSE) can be used as the objective function to be minimized. mathematically SSE is represented as follows:

$$SSE = \sum_{i=1}^k \sum_{x \in c_i} L2(c_i, x)^2$$

L2 is the Euclidean distance between points in the Euclidean space.

one problem of k means is we have to perform multiple runs to choose the most apt solution that is assignments which are frequently into same clusters. This is because of the initialization of centroids issue in k means. we use Hartigan and Wong implementation of k means which is a default method in R statistical tool.

3.4.3 CLUSTER QUALITY EVALUATION: SILHOUETTE INDEX

Another important part of clustering is determining the quality of clusters. For cluster quality determination various methods have been suggested. Two types of cluster quality evaluation metrics are present

Internal evaluation metrics

Internal evaluation metrics are those when we use the same data clustered to verify the quality of the clustering, generally through a score assigned. popular internal evaluation metrics are Silhouette index, Davies Boulden index, Dunn index etc.

External evaluation metrics

External evaluation metrics are those when we have data other than the data used for clustering like bench marks or pre labeled data (human experts). Popular external evaluation metrics are Jaccard index, F measure, Rand measure etc.

In this research, we use silhouette width as cluster quality metric which is further explained as follows:

Silhouette index:

Silhouette index is a function of average distance to points within same cluster and average distance to points in neighboring clusters. Mathematically silhouette index can be represented as follows. Let $a(i)$ be the average dissimilarity of i with points in the same cluster, while $b(i)$ be the average dissimilarity of i with points in neighboring clusters.

silhouette index $s(i)$ can be defined as follows

$s(i) = \frac{a(i) - b(i)}{\max\{a(i), b(i)\}}$. Silhouette index falls in the range of -1 to 1

While silhouette index close to 1 indicates a high cluster, in other words average $s(i)$, the average of silhouette indices of points in a cluster if near to 1 indicates that points are more nearer to each other within cluster and far away from other clusters. if silhouette is close to 0 it implies a point is on border of two clusters or outliers.

we use the silhouette function from R statistical package to compute the silhouette information of a cluster fit/clustering assignment. inputs to the function are clustering assignments of each data point produced by the clustering algorithm used and the second important input is the distance/dissimilarity matrix of the data. This function returns an object of class silhouette in R . silhouette objects are used to generate a silhouette plot which indicates average silhouette width of each cluster and whole assignment.

Chapter 4

RESULTS AND INTERPRETATION

In the results section we initially do inter group participation analysis, which compares the users' in different answering ranges like the sample, low, medium, high and non participators using one way analysis of variance and Tukey HSD statistical tests. In the following section we examine each specific group separately, cluster the users using the feature set and present subtler subgroups of participation in each of the user groups.

4.1 INTER GROUP PARTICIPATION

This section discusses results from comparing all answering groups across each feature. One way analysis of variance has shown that all the 15 attributes are significant in separating the groups. Table 1 presents the F value and *p* value for each feature from one-way ANOVA.

Feature	F value	p value
Questions Asked	29.24	<2e-16
Avg Replies Got	9.547	1.12e-07
Avg Replies Given	828.1	<2e-16
Total Duration	512.8	<2e-16
Active Weeks	2375	<2e-16
Question Words	1151	<2e-16
Answer Words	473.9	<2e-16
Best Participation Point	3130	<2e-16
Best Given	12552	<2e-16

Previous Performance	451.1	<2e-16
After Performance	550.3	<2e-16
Top week	24.82	<2e-16
Performance in top week	1957	<2e-16
Average topics per week	2258	<2e-16
Similarity	286.9	<2e-16

Table 1: F and p values for each feature across user answering ranges from analysis of variance.

In the following section mean and standard deviation of each feature across the groups are presented in a comparative fashion to get a sense of participation inside each group. Further to examine pairwise relationship between groups post hoc analysis is performed using Tukey HSD test.

Questions asked:

Mean and SD of questions asked by users in each answering range

Non participators	Sample participators	Low participators	Medium participators	High participators
2.21 {3.97}	1.82 {4.45}	4.44 {11.26}	8.97 {30.61}	7.75 {15.90}

Users in relatively higher answering ranges asked higher questions but there is no significant difference in question asking behavior of medium and high participators, or non participators and samplers.

Average replies got:

Mean and SD of average replies got for their questions by users in each answering range

Non participators	Sample participators	Low participators	Medium participators	High participators
2.29 {3.13}	1.72 {4.36}	2.63 {4.81}	3.59 {5.13}	4.14 {5.42}

Tukey test showed sample participators are different from all the other groups in getting replies to their questions more on the lower side. Also non participators differed from high participators for the replies they got. All other groups showed no differences.

Average Replies given:

Average replies gives the average of the total number of replies in which a user answers a question. Means and Standard deviations of each subgroup for average replies given as follows

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	1.59 {1.43}	2.01 {1.44}	2.76 {2.88}	2.29 {1.25}

Low participators - high participators and medium participators - high participators did not show significant difference between their means in Tukey test while interestingly medium participators tend to give higher replies on an average.

Total duration and active weeks:

Means and Standard deviations of each subgroup for total duration and active weeks as follows. These two core temporal features are directly proportional to participation and Tukey showed a clear distinction between each group.

Non participators	Sample participators	Low participators	Medium participators	High participators
1.66{5.93}	5.03 {12.03}	19.83 {26.22}	31.38 {26.22}	74.20 {43.13}

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	1.23 {0.56}	3.48 {2.13}	7.58 {3.86}	17.0 {10.68}

Question words and Answer words:

The count of number of words in question content and replies content. Question words were not much different across all answering ranges. In case of answering words there was no significant difference between low-medium and medium-high answering ranges whereas samplers answered in relatively low words compared to other users.

question words

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	23.80 {19.14}	23.98 {9.24}	25.99 {6.78}	26.22 {4.13}

answer words

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	37.53 {50.06}	51.20 {35.25}	59.05 {42.61}	69.96 {43.73}

Best participation point and top week:

Best participation point is the instance in user time period where he answers highest number of questions. Top week attribute is the week number where user answers highest number of questions. Best participation point depends on the top performing week directly and duration inversely. Below are means and SD's of best participation point and top week.

We can see that in lower answering ranges though the top week is early, best participation point is high which is due to low durations given where as it is exactly opposite in higher answering ranges where the best participation points are low though the peak week is high because of relatively higher durations. There is no significant difference in best participation points of low and medium participators.

best participation point table

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	0.81 {0.33}	0.45 {0.37}	0.38 {0.32}	0.17 {0.17}

top week table

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	2.43 {5.46}	5.35 {9.39}	10.08 {13.29}	12.87 {16.44}

Best given and performance in peak week.

Best given is the percentage of answers given in the top week and performance in peak week is the number of answers given in top week. Below are the mean and SD's of these features. Best given is high in lower answering ranges comparatively because of less persistence of users in these groups and low activity periods and vice versa in higher answering ranges where the persistence and activity are high.

best given

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	0.92 {0.18}	0.63 {0.25}	0.45 {0.26}	0.26 {0.13}

performance in top week.

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	1.56 {0.93}	6.19 {3.04}	14.08 {9.68}	33.56 {23.31}

Previous and After performances:

previous performance is the number of questions answered before best participation point/top week, while after number of questions answered after best participation point/top week. On an average for all the groups previous performances are low compared to after performance.

Before performance

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	0.05 {0.27}	1.09 {2.08}	5.60 {6.20}	50.59 {73.63}

There is no significant difference in previous performances of lower 3 answering ranges especially Tukey showed high overlap in non participators and sample participators, this is due to the zero previous participation in samplers.

after performance

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	0.19 {0.52}	2.85 {2.76}	10.67 {8.28}	75.56 {100.19}

After performance of samplers and non participators has high overlap indicating that samplers were mostly one timers mostly and showed no persistence.

Average topics per week:

Average of the count of topics users addressed each week. higher answering ranges addressed higher topics compared to relatively lower user groups on an average.

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	4.73 {3.04}	9.47 {4.77}	10.56 {4.04}	14.06 {3.42}

Similarity of questions asked and answered:

Topical cosine similarity of questions asked and questions answered. There was not much significant difference in topical similarity in answering ranges, This could be due to the fact that low answerers were also low question askers and vice versa. Tukey showed a significant difference only among samplers and low participators.

Non participators	Sample participators	Low participators	Medium participators	High participators
0 {0}	0.27 {0.41}	0.21 {0.33}	0.30 {0.31}	0.26 {0.26}

Post hoc analysis using Tukey HSD test is presented in Appendix.

4.2 INTRA GROUP PARTICIPATION

In this section, we will discuss the sub groups in each of the answering ranges starting with high answerers. The goal of this section is to examine participation differences within answering ranges across the feature set extracted.

4.2.1 HIGH PARTICIPATORS

High participators/answerers are those users who have answered more than 50 questions. This answering range is the widest which contains users who have answered 654 questions. This range is also the smallest but most contributive one.

Clustering results of these high participators revealed 4 low quality clusters. The average silhouette width of multiple k means clustering assignments was between 17 to 20 percentage, which indicates that they are very loosely bound and no strong pattern exists when all the attributes are considered together but thanks to statistical tests which gave some interesting results on few attributes which separated the groups well. These few attribute differences form characteristic of these sub groups.

Before describing each subgroup, one way ANOVA shows that for the subgroups replies given, question and answer words, answers in peak week do not vary significantly. The table highlights these attributes and values.

Questions Asked	14.68	6.21e-06
Avg Replies Got	12.07	3e-05
Avg Replies Given	0.815	0.496
Total Duration	5.353	0.00484
Active Weeks	9.276	0.000201
Question Words	2.025	0.133

Answer Words	2.214	0.109
Best Participation Point	22.14	1.5e-07
Best Given	13.64	1.14e-05
Previous Performance	4.011	0.0171
After Performance	3.109	0.0423
Top week	40.1	2.87e-10
Performance in top week	0.644	0.593
Average topics per week	3.8	0.021
similarity	5.105	0.00607

Table 2: *F* and *p* values for each feature across high answerers from analysis of variance.

User subgroups in high answerers are

Group 1: High answerers with short life and less consistency

They asked less questions, stayed for lower durations and were active for very less time compared to other high answerers. They peaked/performed best in their initial weeks and gradually declined in participation. These users answered 40 percent of their answers in peak week. Additionally, their average topics per week is relatively highest among all user groups. In the following mean and SD tables we highlight the key features which reflect the group nature.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
4.60{5.08}	3.01{3.3}	2.11{0.79}	40.07{26.38}	6.90{2.94}	28.67{4.46}	72.5{38.02}

B.P.P: best participation point B.G: best given PWP : peak week performance

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.07{0.07}	0.41{0.11}	1.5{0.67}	37.2{12.85}	10.9{16.70}	49.4{16.74}	16.74{2.65}	0.38{0.32}

Group 2: Consistent high answerers and early performers

Low question askers. They stayed for longer durations and activity period is high as well. They too peaked initially as we can see from their best participation point but were quite consistent as from their previous and after performances. Their top 3 weeks of participation were 5-10 weeks which implies their major participation was in early stages.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
2.61{3.58}	2.98{4.35}	2.04{0.78}	84.26{35.42}	20.30{6.46}	25.45{3.46}	56.34{23.44}

B.P.P: best participation point B.G: best given PWP : peak week performance

B.P.P	B.G	peak week	PWP	Previous performance	after performance	topics per week	similarity
-------	-----	-----------	-----	----------------------	-------------------	-----------------	------------

0.08{0.08}	0.19{0.05}	5.53{4.53}	36.61{31.46}	38.30{38.20}	134.00{133.15}	13.00{3.14}	0.18{0.18}
------------	------------	------------	--------------	--------------	----------------	-------------	------------

Group 3: Consistent high answerers and later performers

Low question askers with highest durations in the group, their activity period was also high. These users peaked comparatively later in their lifetime. These high answerers were consistent through their lifetime in the community and had high previous performance in exact contrast to group 2 users who had high after performance and low previous performance. Their top 3 weeks of participation were 30-40 weeks which implies their major participation was in later stages.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
1.2{2.4}	0.96{1.93}	2.83{2.49}	116{43.03}	18.4{10.65}	25.3{3.80}	111.19{72.82}

B.P.P: best participation point B.G: best given PWP : peak week performance

B.P.P	B.G	peak week	PWP	Previous performance	after performance	topics per week	similarity
0.38{0.12}	0.24{0.09}	40{11.36}	29.6{18.35}	112.6{112.4}	8.8{8.42}	13.15{3.62}	0.01{0.02}

Group 4: High question askers who were consistent and showed higher similarities

These users were high question askers and produced high conversations in their questions, which is their most distinctive feature among other high answerers. They like other high answerers were active for long and stayed in the community for higher durations. They peaked in the later stage of their participation, they were consistent in participation

like other high answerers. Though the average topics they addressed is low, their similarities between questions asked and answered is highest.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
40.5 {25.93}	14.74 {2.98}	2.9 {0.51}	74.6 {36.42}	29.75 {12.61}	23.65 {2.34}	56.2 {18.35}

B.P.P : best participation point B.G: best given PWP : peak week performance

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.4 {0.12}	0.14 {0.00}	31.2 {13.38}	19.5 {4.09}	112.2 {8.38}	34.5 {24.45}	11.95 {0.70}	0.52{0.09}

Post hoc analysis using Tukey HSD test is presented in Appendix.

4.2.2 MEDIUM PARTICIPATORS

Users who have answered in the range of (20,50] are medium participators. we found total 46 users in this range. Like high answerers the cluster quality was low (silhouette width around 20 percent for multiple k means iterations) but statistical tests has shown some interesting participation patterns within medium participators. Total 3 sub clusters are studied

Before we discuss each subgroup we present the one ANOVA results of each feature for the clustered populations. All the highlighted features, that is the questions asked, replies given, replies got, duration, question words, answer words, similarities had no significant difference in separating the groups. We can observe that mostly temporal features played part in distinguishing these groups.

Feature	F value	p value
Questions Asked	1.68	0.198
Avg Replies Got	1.609	0.212
Avg Replies Given	0.233	0.793
Total Duration	1.517	0.231
Active Weeks	17.03	3.57e-06
Question Words	2.635	0.0833
Answer Words	2.225	0.12
Best Participation Point	16.22	5.64e-06
Best Given	44.81	3.05e-11
Previous Performance	40.44	1.32e-10
After Performance	22.94	1.66e-07
Top week	32.64	2.38e-09
Performance in top week	65.12	9.75e-14
Average topics per week	24.18	9.2e-08
similarity	1.222	0.305

Table 3: F and p values for each feature across medium participators from analysis of variance

Group1: Late Medium answerers

These Users stayed active for high time. They peaked later in their duration , gave around one third percent of their answers in the peak week. Their top 3 weeks of participation were 10-20 weeks which implies their major participation was in later stages.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
10.38 {11.38}	4.57 {4.68}	2.62 {2.35}	41.35 {22.64}	9.46 {3.75}	25.9 {5.48}	50.01 {24.55}

B.P.P : best participation point B.G: best given PWP : peak week performance

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.70 {0.25}	0.36 {0.17}	26.38 {13.9}	10.00 {4.88}	13.30 {4.26}	4.53 {3.65}	8.44 {2.11}	0.42 {0.31}

Group 2: Early medium answerers

These group of users are similar to group 1 of users in most features but show a contrast when it comes to peaking. They have peaked in initial stages in the community. They showed high after performance unlike high previous performance of group 1 users. Their top 3 weeks of participation were 1-6 weeks which implies their major participation was in early stages.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
2.5 {5.22}	4.08 {5.79}	3.02 {3.44}	30.40 {30.24}	8.50 {2.78}	24.41 {6.05}	71.11 {50.45}

B.P.P: best participation point B.G: best given PWP: peak week performance

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.21 {0.16}	0.33 {0.14}	4.66 {5.45}	9.91 {4.30}	3.54 {3.96}	16.37 {6.80}	9.49 {2.51}	0.24 {0.23}

Group 3: Least active medium participators

These group of users were least active among the medium participators. They peaked at 40 percent of their journey but gave 90 percent of their answers in a one shot time period that too in their first week, also strengthened by zero previous performance. Their after performance was too low suggesting one time performance and a quick dilution in participation. Striking thing is they dealt very high topics in the range of high answerers.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
24.22 {64.68}	0.88 {1.26}	2.26 {1.55}	19.58 {29.34}	2.44 {1.25}	30.35 {8.26}	39.95 {27.43}

B.P.P: best participation point B.G: best given PWP: peak week performance

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.39 {0.36}	0.89 {0.12}	1.00	31.11 {5.92}	0	4.33 {5.20}	16.46 {4.01}	0.29 {0.42}

Post hoc analysis using Tukey HSD test is presented in Appendix.

4.2.3 LOW PARTICIPATORS

Low participators are those users who have answered more than 5 questions and at most 20 questions. Clustering low participators gave 5 sub groups. Anova across these subgroups showed that replies given, question words and answer words have no significance and are not predictors of sub groups.

Feature	F value	p value
---------	---------	---------

Questions Asked	8.254 3.84e-06
Avg Replies Got	24.36 3.58e-16
Avg Replies Given	0.6 0.66
Total Duration	10.94 5.61e-08
Active Weeks	66.34 2e-16
Question Words	1.624 0.17
Answer Words	1.033 0.391
Best Participation Point	156.1 2e-16
Best Given	113 2e-16
Previous Performance	41.18 2e-16
After Performance	54.58 2e-16
Top week	40.48 2e-16
Performance in top week	23.75 7.77e-16
Average topics per week	59.06 2e-16
Similarity	161.6 2e-16

Table 4: *F* and *p* values for each feature across low participators from analysis of variance

Group 1 (low participators with high similarity)

These users asked high questions and got high replies. They were active were low periods and gave their $\frac{3}{4}$ th of best in one time unit with very low previous and after performances. Interesting feature is high similarity of questions asked and answered. These users have given low percentage of their duration in answering questions.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
12.08	7.69	2.12	15.70	2.79	23.6	46.44

{14.34}	{7.06}	{1.29}	{24.6}	{1.40}	{9.6}	{45.22}
---------	--------	--------	--------	--------	-------	---------

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.34 {0.29}	0.69 {0.19}	2.50 {3.69}	6.29 {2.39}	0.52 {1.00}	2.5 {1.95}	9.77 {3.60}	0.83 {0.16}

Group2 (low participators who addressed high topics and least active):

They asked very low questions, peaked in the end and gave 90 percent of their best and have least active weeks and duration. They were highly enthusiastic in the week they contributed and never came back.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
0.03 {0.17}	0.13 {0.71}	1.98 {1.90}	0.64 {0.69}	1.06 {0.24}	25.99 {10.99}	56.05 {35.85}

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.98 {0.05}	0.98 {0.03}	1.13 {0.56}	8.56 {2.70}	0.03 {0.17}	0.03 {0.17}	16.71 {4.81}	0

Group 3 (low participators with longest life):

These users are low participators who stayed for long durations, asked high questions got good replies, they were also highly active in terms of active weeks, they

peaked initially, quite sustained compared to others. Irrespective of high durations their contribution was low indicating high gap in their journey.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
7.33 {17.17}	4.53 {5.07}	1.97 {0.76}	38.91 {25.3}	6.27 {2.02}	21.42 {7.75}	49.72 {21.46}

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.14 {0.16}	0.31 {0.09}	5.5 {7.23}	3.5 {1.67}	1.2 {1.90}	6.3 {2.25}	5.32 {1.91}	0.2 {0.24}

Group 4 (laziest low participators):

These users asked low number of questions, got low replies, addressed medium topics. But they are too lazy as they stayed for high durations in the range of 20 weeks with low active period of 2 weeks. So they have not used their duration for either asking or answering questions which implies this time they might have stayed out of community.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
0.82 {3.66}	0.44 {1.19}	1.83 {1.28}	19.10 {29.9}	2.0 {1}	24.48 {8.79}	56.83 {34.63}

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.22 {0.16}	0.68 {0.14}	1.68 {3.86}	7.43 {3.08}	0.22 {0.76}	3.22 {2.08}	9.50 {3.01}	0.018 {0.07}

Group 5 (late low participators):

These users are medium question askers. They stayed for long durations with decent active life. They peaked much later in their life at almost end. They answered 50 percent of questions in the peak week. Their 3 peak weeks were in the end of their journey and give the long journey, they are late in answering.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
3.73 {7.76}	1.38 {1.98}	2.31 {1.88}	22.25 {16.77}	4.3 {1.57}	22.59 {8.38}	49.19 {33.8}

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.87 {0.13}	0.48 {0.13}	19.73 {13.64}	4.53 {1.64}	4.23 {2.75}	1.13 {1.72}	6.78 {2.85}	0.12 {0.21}

Post hoc analysis using Tukey HSD test is presented in Appendix.

4.2.4 SAMPLE PARTICIPATORS

In sample participators replies given, question words and answer words have no significance and are not predictors of sub groups. Their typical active answering time is 1-2 weeks. 4 sub sample participators have been found which are explained as follows

Group 1 (one time samplers with high similarity)

They asked good number of questions for the group. They answer very low (1-2) in peak week but similarity is 0.9 which says they answer relative to what they asked, implying more similar questions recommended would have had them stick.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
3.03 {4.96}	3.46 {5.26}	1.45 {1.33}	2.72 {8.01}	1.02 {0.15}	21.49 {15.68}	33.40 {40.79}

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.91 {0.21}	0.99 {0.056}	2.61 {6.26}	1.33 {0.55}	0.01 {0.12}	0.008 {0.09}	4.03 {2.44}	0.94 {0.118}

Group 2 (long livers with high inactivity):

They ask good number of questions and get good replies, stay for longer durations , peak initially and give 70 percent of their best(though they answer 1-2 questions), most of the time is non participative.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
4.01 {6.60}	2.80 {5.10}	1.70 {1.45}	17.03 {18.61}	1.92 {0.79}	23.36 {15.48}	34.86 {35.97}

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.21 {0.18}	0.66 {0.24}	2.41 {5.00}	1.43 {0.62}	0.05 {0.23}	0.90 {0.82}	4.10 {1.95}	0.21 {0.30}

Group 3 (one time samplers who address high topics):

They ask lower questions comparatively, stay for medium time period (on average 5-6 weeks), they answer relatively high questions (3-4) and addressed higher topics in the one time unit.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
1.07 {2.59}	1.25 {3.55}	1.58 {0.95}	5.06 {12.29}	1.39 {0.69}	25.52 {13.66}	44.73 {38.01}

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.88 {0.25}	0.90 {0.16}	3.81 {7.99}	3.49 {0.89}	0.27 {0.62}	0.14 {0.39}	9.14 {3.58}	0.19 {0.34}

Group 4 (typical samplers):

lowest questions (0.5) asked, most of them answer one question and never come back.

Questions asked	Replies got	Replies given	Total Duration	Active weeks	Question words	Answer Words
0.54 {2.62}	0.58 {3.25}	1.60 {1.56}	1.16 {4.55}	1	24.59 {22.87}	38.58 {60.68}

B.P.P	B.G	peak week	PWP	previous performance	after performance	topics per week	similarity
0.98 {0.069}	1	1.97 {4.42}	1.19 {0.39}	0	0	4.08 {2.40}	0.009 {0.05}

ANOVA results

Feature	F value	p value
Questions Asked	42.87	2e-16
Avg Replies Got	31.82	2e-16
Avg Replies Given	1.23	0.297
Total Duration	120	2e-16
Active Weeks	253.7	2e-16
Question Words	1.898	0.128
Answer Words	1.825	0.141
Best Participation Point	1129	2e-16
Best Given	393.1	2e-16
Previous Performance	46.26	2e-16
After Performance	296.6	2e-16
Top week	4.236	0.00548
Performance in top week	644.3	2e-16
Average topics per week	168.5	2e-16
Similarity	1297	2e-16

Table 5: F and p values for each feature across sample participators from analysis of variance

Post hoc analysis using Tukey HSD test is presented in Appendix.

4.2.5 NON PARTICIPATORS

This category had only 3 valid attributes that is the questions asked, replies got and total duration. The number of sub clusters are 6 and all three attributes could differentiate well in terms of clustering

Groups 1,2 (one timers): Two sub groups were one timers with low durations [<1 week] and asked around 1 question with little bit variance in replies got

Mean[ques_asked]	STD[ques_asjed]	Mean[avg_replies_got]	STD[avg_rep lies_got]	Mean[total_duration]	STD[total_d uration]
1.42	1.00	0.65	0.51	0.26	0.80

Mean[ques_asked]	STD[ques_asjed]	Mean[avg_replies_got]	STD[avg_rep lies_got]	Mean[total_duration]	STD[total_d uration]
1.45	0.98	3.30	1.53	0.31	0.86

Group 3 (high question askers):

These users were high question askers and stayed longer durations with around 3 replies got an average.

Mean[ques_asked]	STD[ques_asjed]	Mean[avg_replies_got]	STD[avg_rep lies_got]	Mean[total_duration]	STD[total_d uration]
37.55	12.57	2.82	1.38	16.06	12.05

Group 4 (highly conversationalists)

These users asked less questions but produced very high conversations on their questions.

Mean[ques_asked]	STD[ques_asjed]	Mean[avg_replies_got]	STD[avg_rep lies_got]	Mean[total_duration]	STD[total_d uration]
2.07	2.63	14.55	7.29	1.01	2.61

Group 5,6 (Lazy participators) :

There are two variations of these users who are highly lazy and relatively less lazy. Both these categories of users stayed longer durations on average (38,9) but asked 6-7 questions on average.

Mean[ques_asked]	STD[ques_asked]	Mean[avg_replies_got]	STD[avg_replies_got]	Mean[total_duration]	STD[total_duration]
6.08	4.97	3.83	3.69	38.14	14.74

Mean[ques_asked]	STD[ques_asked]	Mean[avg_replies_got]	STD[avg_replies_got]	Mean[total_duration]
7.03	4.65	2.63	2.00	9.20

ANOVA results:

attributes	F value	p
Question Asked	1350	<2e-16
Replies Got	1010	<2e-16
Total Duration	1705	<2e-16

Table 6: *F* and *p* values for each feature across Non Participators from analysis of variance

Post hoc analysis using Tukey HSD test is presented in Appendix.

Summary of results: Important patterns(properties) and corresponding groups

[Relative observations]

Pattern	HP	MP	LP	SP	NP
Initial Enthusiasts					
Early Participators					
Late Participators					
High askers and answerers					
High question askers					
High conversationalists					
One timers					

Lazy participators						
--------------------	--	--	--	--	--	--

4.2.6 PATTERNS AND INTERVENTIONS

Pattern: Initial Enthusiasts

User patterns found in high and medium participators show a group of users who are initial enthusiasts. By initial enthusiasm, we mean that they have answered very high questions in single week (time unit) and could not sustain participation at all later. Participation in rest of the active time is either very low or no where comparable to their contribution in the initial weeks.

This initial enthusiasm if sustained over the year, can improve overall participation. Similar patterns were found in low answering ranges making initial enthusiasm an important pattern which needs attention from community designers.

This issue can be possibly dealt with some measures like motivating user with new user titles which reflect their swiftness and enthusiasm positively. This creates a status for such users in the community. For those users who never come back, sending reminders and recommending similar questions could help them to come back and contribute.

Pattern: Experts who are also learners.

As seen from high answerers, most of the subgroups ask very little questions or none at all. These users are more focused on answering more. They tend to be only contributors. But, another pattern found in high answerers was that of users who asked high questions which is not typical of high answerers. This implies that these users have used the community to gain information from peers while contributing highly to the community

by answering in a sustained manner for long durations. These users are not typical experts and also learning oriented.

Routing the questions these users ask to the typical experts already existing might result in high quality discussions. This is good for both categories of users and for the community to create a vibrant expert base. Special titles given and correspondingly giving higher scores to these users than typical experts could motivate more users to actually use the community for learning.

Pattern: Early and Late participators

Some patterns found in medium and high participators show groups of users who are either early participators or late participators. These users either perform maximum of their best in initial part of their life or later part. This implies that at least half of their time has not been used well. The durations and active weeks of these users were pretty high attesting the pattern.

This pattern of participation also needs the attention of community designers. These groups of users who have participated pretty steadily should be motivated to participate and contribute in their less active periods by some measures like decreasing scores, dropping titles. These activities reflect on their social fame and might get their attention.

Patterns in lower answering ranges

Higher conversationalists:

These users who produced very high discussions on few questions they asked remained as low participators. Their high conversational nature indicates their interest as well as their learning oriented behavior. These users prefer to get their questions answered in detail rather getting answers and leaving the community.

Community designers should target such users and motivate them to ask more questions and also suggestions/recommendations to answer questions similar to the topics in which they asked.

High topic addressers:

These low answering users have interestingly addressed very high topics in the range of high answerers but could not sustain. These users could be experts who did not sustain in the community.

Motivating such users could turn them to high contributors. These users could be identified in the initial stages by looking at their answering nature. Special titles, improving smart scores might motivate them to sustain in the community.

4.2.7 KNOWN LIMITATIONS

The participation patterns found cannot be generalized to all of OpenStudy. This is because the study was performed only on single chemistry subject group. There are groups like physics and mathematics which are comparatively larger in size. This implies more number of users, especially in higher answering ranges. So the patterns found out might not be complete to describe users in all categories in all groups.

Participation itself is a very complex variable. It can be measured in several dimensions. In this study participation analysis can't be a complete analysis owing to the fact that several unexplored dimensions exist like the quality of content contributed because it is very subjective and demands a comprehensive study which includes text, contextual and emotional aspects. Social networking aspect of participation is also unexplored which might involve identification of influential communities of interaction, frequent collaborators etc.

Personal messages were not open and a private content. Some part of personal messages including group chats might include some new participation patterns which are hidden like the informal interactions using internet emoticons, smileys etc.

This study could not make full use of the already existing participation/contribution metrics in OpenStudy for various reasons. The smart score of users currently in use is the combination of problem solving, team building and engagement scores. Since the details of calculation are abstract, that is we do not know what algorithms/factors decided problem solving exactly we did not consider to use them or to evaluate users based on their smart scores. Another reason we did not consider is we were not sure whether the initial database users smart scores and much recent ones are on a same scale.

Another OpenStudy metric we decided not to use to analyze participation was number of medals a user has. This is because it is not clear when, where and how a user can be granted a medal. We found instances where users actually offered medals to get their questions answered in their home works. Interestingly some real well answered replies did not receive medals. This was another limitation where we could not make use of an existing metric.

Few other metrics where we had similar issues of clarity were fans and titles. Users were offered being fans to them if they get their questions answered. User titles, though were interesting, it is hard to compare these titles and gauge the degree of all titles exactly. So, both these metrics were not used for evaluation purposes and also as features in our study.

In this study, weeks were used as time units to understand participation by treating each of them single unit. This decision has been made by considering that week's time is

substantial to measure user's instance of interaction with community. But it would be interesting to further explore how users participation vary within each week. This might give new views into participation.

4.2.8 CONCLUSIONS AND FUTURE WORK

To the best of our knowledge, past work in question and answer communities has not explored user participation across broad spectrum of attributes like it was explored in this thesis. Content specific dimensions like count of question words, answer words, average topics addressed in a week and similarity of questions asked and questions answered were addressed. Majority of studies in this area used topic modelling to recommend similar questions for answering, searching most relevant questions to user input and assess expertise. In this study we used to test the similarity in questions asked and questions answered to understand whether similarity might imply learning behavior. We also compared user groups and subgroups based on the count of topics they addressed in a week on average. Count of question words and answer words was used to understand whether length /size of the content is related to participation. We also explored temporal features in a new manner to figure out initial enthusiasm and sustenance in participation with the help of total duration, number of active weeks, top performing point/week, performance percentage in the top week(best given), previous and after performances. We also used questions specific participation attributes like questions asked, average replies got and average replies given. Previous work in Yahoo answers examined user participation across what yahoo calls maturity levels [10]. Few quality of participation metrics like percentage of answers which received stars, percentage of best answers were also analyzed. In this study, user levels were defined and various participation features

were analyzed across these groups to get a better sense of each user types discovered. Past work has also shown top angle/external views do not give a complete picture. For example MOOCs engagement patterns discovered with the help of clustering engagement labels in a timeline [11]. We were also motivated in similar lines , to get a better view of participation by exploring participation features across multiple dimensions and ended up studying various categories of interesting user patterns in an open question and answer based community. Like quite a few studies which used cluster analysis to throw light on hidden patterns in educational settings [15-17,19,21], we used K means clustering technique in user groups to find hidden sub patterns to get a deeper understanding of participation. While some the literature has explored more than one clustering algorithm to compare the efficiency, this study focused on single most popular K means algorithm as we concentrated more on comparing the user groups obtained. In future, there is a broad scope to test various classes of clustering algorithms to further expand understanding by obtaining better clusters. We found some interesting participation patterns which need the attention of community designers to help the community and users to sustain. We tried to discuss certain patterns and interventions possible like it was done in engagement pattern study in MOOCs.

We have considered only chemistry community from OpenStudy which is comparatively less in size. In future, we will try to explore communities like mathematics and physics, which have larger user base. We will try to verify if similar user participation patterns are also found in other communities.

We used k means algorithm to cluster users and find any hidden patterns. It is well studied in literature but has some inherent limitations in clustering complex data, especially

multi dimensional and non spherical clusters. We plan to use a better clustering algorithm like Expectation Maximization or Spectral clustering, which are getting more popular in this domain. Given more time we could have done a comparative study of clustering algorithms to evaluate the best algorithm which works in this work.

We also plan to expand on topic modelling process to form user models which store the topic variation with time, improvement in topic strengths of the user with time, most active topics. These information tells more about users and helps in getting deeper insights into learning processes in such communities. Topic evaluation by experts is another task which we could have done given more time.

Despite some limitations in the study, given the broad scope of future research, especially in new subject groups, expanding on the clustering techniques and feature dimensions, we believe this work is a big step in understanding participation in question and answer communities, especially those with learning implications.

REFERENCES

- [1] McLoughlin, C., & Lee, M. J. (2007, December). Social software and participatory learning: Pedagogical choices with technology affordances in the Web 2.0 era. In *ICT: Providing choices for learners and learning. Proceedings ascilite Singapore 2007* (pp. 664-675).
- [2] Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008, April). Predictors of answer quality in online Q&A sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 865-874). ACM.
- [3] Yahoo Answers Hits 300 Million Questions, But Q&A Activity Is Declining. (2012, July 10). Retrieved October 28, 2015.
- [4] Ram, A., Ai, H., Ram, P., & Sahay, S. (2011, May). Open social learning communities. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics* (p. 2). ACM.
- [5] OpenStudy. (n.d.). Retrieved October 28, 2015, from <http://openstudy.com/code-of-conduct>
- [6] Gazan, R. (2011). Social q&a. *Journal of the American Society for Information Science and Technology*, 62(12), 2301-2312.
- [7] Lueg, C., & Fisher, D. (Eds.). (2012). *From Usenet to CoWebs: interacting with social information spaces*. Springer Science & Business Media.
- [8] Points & Levels | Yahoo Answers. (n.d.). Retrieved November 6, 2015, from https://answers.yahoo.com/info/scoring_system
- [9] What is reputation? How do I earn (and lose) it? - Help Center. (n.d.). Retrieved November 6, 2015, from <http://stackoverflow.com/help/whats-reputation>
- [10] Shah, C., Oh, J. S., & Oh, S. (2008). Exploring characteristics and effects of user participation in online social Q&A sites. *First Monday*, 13(9).
- [11] Kizilcec, R. F., Piech, C., & Schneider, E. (2013, April). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170-179). ACM.
- [12] Educational Data Mining. (n.d.). Retrieved November 6, 2015, from <http://www.educationaldatamining.org/>

- [13] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6), 601-618.
- [14] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1), 3-17.
- [15] Amershi, S., & Conati, C. (2007, January). Unsupervised and supervised machine learning in user modeling for intelligent learning environments. In *Proceedings of the 12th international conference on Intelligent user interfaces*(pp. 72-81). ACM.
- [16] Trivedi, S., Pardos, Z., Sárközy, G., & Heffernan, N. (2010, June). Spectral clustering in educational data mining. In *Educational Data Mining 2011*.
- [17] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2, 849-856.
- [18] Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.
- [19] Xu, B. (2011). Clustering educational digital library usage data: Comparisons of latent class analysis and K-Means algorithms.
- [20] Exploratory Learning. (n.d.). Retrieved November 6, 2015, from http://www.lynnevey.com/eleweb/index_ele.html
- [21] Amershi, S., & Conati, C. (2010). Automatic recognition of learner types in exploratory learning environments. *Handbook of educational data mining*, 213.
- [22] Free Online Courses - FutureLearn. (n.d.). Retrieved November 6, 2015, from <https://www.futurelearn.com/>
- [23] Ferguson, R., & Clow, D. (2015, March). Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 51-58). ACM.
- [24] Chodorow, K. (2013). *MongoDB: the definitive guide*. " O'Reilly Media, Inc."
- [25] BSON. (n.d.). Retrieved November 6, 2015, from <http://docs.mongodb.org/meta-driver/legacy/bson/>
- [26] JSON and BSON. (n.d.). Retrieved November 6, 2015, from <https://www.mongodb.com/json-and-bson>
- [27] Tan, P. N., Steinbach, M., & Kumar, V. (2013). *Data Mining Cluster Analysis: Basic Concepts and Algorithms*.

[28] Topic Modeling. (n.d.). Retrieved November 23, 2015, from <http://mallet.cs.umass.edu/topics.php>

APPENDIX A
POST HOC TESTING RESULTS

In this section we list Post hoc testing results of all the results discussed in the results and interpretation section.

A.1 POST HOC TESTING USING TUKEY HSD FOR INTER GROUP

PARTICIPATION

We use the following short form notations in this section,

For users:

NP = Non Participator

SP = Sample Participator

LP = Low Participator

MP = Medium Participator

HP = High Participator

	LP-HP	MP-HP	NP-HP	SP-HP	MP-LP	NP-LP	SP-LP	NP-MP	SP-MP	SP-NP
Questions Asked	0.03	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.36
Avg Replies Got	0.20	0.96	0.03	0.00	0.50	0.72	0.01	0.11	0.00	0.00
Avg Replies Given	0.46	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total Duration	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Active Weeks	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Question Words	0.80	0.99	0.00	0.70	0.77	0.00	0.99	0.00	0.63	0.00
Answer Words	0.00	0.46	0.00	0.00	0.45	0.00	0.00	0.00	0.00	0.00
Best Participation Point	0.00	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.00

Best Given	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Previous Performance	0.00	0.00	0.00	0.00	0.00	0.20	0.29	0.00	0.00	0.99
After Performance	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97
Top week	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Performance in top week	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average topics per week	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
similarity	0.83	0.91	0.0	0.99	0.11	0.00	0.00	0.00	0.89	0.00

Table 7: Tukey HSD test results for all answering ranges

A.2 POST HOC TESTING USING TUKEY HSD FOR INTER GROUP

PARTICIPATION

For subgroups in High Participators

	c2-c1	c3-c1	c4-c1	c3-c2	c4-c2	c4-c3
Questions Asked	0.97	0.93	0.00	0.99	0.00	0.00
Average Replies got	0.99	0.76	0.00	0.74	0.00	0.00
Total Duration	0.03	0.00	0.40	0.37	0.96	0.35
Active Weeks	0.00	0.06	0.00	0.96	0.19	0.18
Best Participation point	0.98	0.00	0.00	0.00	0.00	0.85
Best Given	0.00	0.01	0.00	0.66	0.75	0.31
Previous Performance	0.75	0.04	0.06	0.16	0.22	0.99
After Performance	0.15	0.85	0.99	0.07	0.26	0.97
Top week	0.59	0.00	0.00	0.00	0.00	0.33

Performance in top week	0.03	0.16	0.06	0.99	0.93	0.93
Similarity	0.19	0.03	0.73	0.49	0.06	0.01

Table 8: Tukey HSD test results for subgroups in High Participators

For subgroups in Medium Participators

	c2-c1	c3-c1	c3-c2
Active Weeks	0.62	0.00	0.00
Best Participation Point	0.00	0.01	0.17
Best Given	0.91	0.00	0.00
Previous Performance	0.00	0.00	0.05
After Performance	0.00	0.99	0.00
Top Week	0.00	0.00	0.52
Performance in top week	0.99	0.00	0.00
Average topics per week	0.54	0.00	0.00

Table 9: Tukey HSD test results for subgroups in Medium Participators

For subgroups in Low Participators

	c2-c1	c3-c1	c4-c1	c5-c1	c3-c2	c4-c2	c5-c2	c4-c3	c5-c3	c5-c4
Questions Asked	0.00	0.00	0.00	0.32	0.73	0.99	0.03	0.65	0.63	0.04
Average Replies got	0.00	0.00	0.00	0.00	0.83	0.99	0.00	0.73	0.01	0.00
Total Duration	0.96	0.80	0.09	0.00	0.97	0.00	0.00	0.00	0.04	0.00
Active Weeks	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Best Participation point	0.01	0.00	0.00	0.00	0.00	0.00	0.31	0.13	0.00	0.00
Best Given	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Previous Performance	0.88	0.00	0.69	0.24	0.00	0.98	0.01	0.00	0.00	0.01
After Performance	0.43	0.02	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00
Top week	0.98	0.00	0.93	0.37	0.00	0.99	0.07	0.00	0.00	0.08
Performance in top week	0.22	0.04	0.00	0.00	0.00	0.26	0.00	0.00	0.48	0.00
Average topics per week	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.35	0.00
Similarity	0.00	0.00	0.00	0.00	0.02	0.98	0.00	0.02	0.35	0.00

Table 10: Tukey HSD test results for subgroups in Low Participators

For subgroups in Sample Participators

Post hoc analysis using Tukey HSD

	c2-c1	c3-c1	c4-c1	c3-c2	c4-c2	c4-c3
Questions Asked	0.06	0.00	0.00	0.00	0.00	0.54
Avg Replies Got	0.34	0.00	0.00	0.00	0.00	0.32
Avg Replies Given	0.24	0.80	0.49	0.87	0.84	0.99
Total Duration	0.00	0.14	0.22	0.00	0.00	0.00
Active Weeks	0.00	0.00	0.88	0.00	0.00	0.00
Question Words	0.72	0.18	0.16	0.72	0.86	0.95
Answer Words	0.98	0.13	0.54	0.26	0.79	0.56
Best Participation Point	0.00	0.38	0.00	0.00	0.00	0.00

Best Given	0.00	0.00	0.80	0.00	0.00	0.00
Previous Performance	0.33	0.00	0.84	0.00	0.00	0.20
After Performance	0.00	0.00	0.99	0.00	0.00	0.00
Top week	0.97	0.16	0.44	0.08	0.77	0.00
Performance in top week	0.2	0.00	0.01	0.00	0.00	0.00
Average topics per week	0.99	0.00	0.99	0.00	0.99	0.00
similarity	0.00	0.00	0.00	0.85	0.00	0.00

Table 11: Tukey HSD test results for subgroups in Sample Participators

For subgroups in Non Participators

	Questions asked	Average Replies got	Total Duration
c2-c1	0.00	0.00	0.00
c3-c1	0.00	0.00	0.00
c4-c1	0.00	0.00	0.00
c5-c1	0.06	0.00	0.18
c6-c1	0.99	0.00	0.99
c3-c2	0.00	0.99	0.00
c4-c2	0.00	0.37	0.00
c5-c2	0.00	0.00	0.00
c6-c2	0.00	0.87	0.00
c4-c3	0.99	0.00	0.00
c5-c3	0.00	0.00	0.00
c6-c3	0.00	0.00	0.00
c5-c4	0.00	0.00	0.00
c6-c4	0.00	0.49	0.00

c6-c5	0.10	0.00	0.27
-------	------	------	------

Table 12: Tukey HSD test results for subgroups in Non Participators