DataVis Process Book

Class Brüss and Maximilian Rünz

December 2017

1 Introduction

Stack Overflow[link] has grown to one of the most popular websites not only for computer scientists, but also for the public of coders. 50 Million people per month visit Stack Overflow each month. Since it has become a very important tool for developers we were curious about how the community looks like. With the gained insights it might be possible to make Stack Overflow even more beneficial as it is nowadays. Users can identify their spot in the community in order to help other users individually.

Therefore we have asked ourselves the following questions:

- What types of users to exist? How do they behave?
- Who keeps Stack Overflow running?
- How did Stack Overflow grew to this community?

To answer these questions our analysis and visualization will highlight several facts about the community as it is today and how Stack Overflow with its community growing to such an influencing website. We will highlight discrepancies between users. Based on these observed discrepancies we will classify users into different user types. Furthermore, the behavior of users of the different types will be shown. In the second part of our analysis we will then focus on the evolution of Stack Overflow. We will show how contribution of users changed within the last ten years. And last but not least we will show how users evolved in this time frame.

This visualization mainly addresses Stack Overflow users and aims to give them an impression off the engagement within the larger community and how it evolved over the years. In addition we wanted to allow every Stack Overflow user visiting our site to put themselves into context with the larger engaged community and clearly see how much work is put in by the users that really carry the community.

2 Related Work and Inspiration

The idea of analyzing Stack Overflow came up since during our daily life we discovered insight statistics from Stack Overflow[https://insights.stackoverflow.com/survey/2017]. We realized how much we use the site on a daily basis, but how little we knew about its community structure and the varity of users on the platform. Since we were also thinking about visualizing communication patters of humans we decided to analyse the community of Stack Overflow.

During the creation of our visualization we came across several inspirational visualizations. We liked the story telling of this[http://ncase.me/trust/] post using an iterative and interactive approach. Therefore we decided to guide the user step by step through the visualization of Stack Overflow. The further the user gets in the story the more interactive the visualization gets.

For the visualizations as such we found some inspiring posts, too. Gapminder[https://www.gapminder.org/tools/] shows a very time evolution of three quantative values and one categorial value.

(TODO more?)

3 Dataset and Preprocessing

The analysis and visualizations are based on a data dump from Stack Exchange[link] posetd on the archive[https://archive.org/details/stackexchange]. It is updated frequently, we have used a version from October 2017, but the analysis can easily be redone with a more recent snapshot. We have used the uploaded files for users, tags and posts for the golbal stackoverflow.com. The files are stored in XML format and add up to a size of approx. 100GB.

In order to handle the data we removed for us not necessary information. From the user.xml file we extracted the following features into a JSON format: UserId, DisplayName, CreationDate and Reputation. The post.xml file was split up into two JSON files, one for questions the other one for answers. Both files contain the ID of the user who created the post, its creation date, the number of votes and the linked answer/question and the tags of the question.

In the next step of our prepossessing we created a network view of the community. Therefore, we created based on the questions and linked answers an edge list annotated with the time stamp of the answer and a list of tags for the questions. During the explanatory analysis we noticed that is in-feasible to analyze the community only by looking at the created network of users. We decided to aggregate our reduced information per user. With the already reduced Question and Answer data we created a summary for each user over time. The resulting JSON file includes the user ID, Reputation, creation date and name as general features. Furthermore, it contains for each user per year the number of questions and answers posted, the average votes for questions and answers and the number of votes of all questions and answers posted.

Table 1: Network metrics JavaScript community

Feature	# Nodes
# Nodes	417219
# Edges	765669
# Conn. Components	36140
Avg. Node Degree	3.67
Avg. Clustering	0.0014
Giant Component	357624

4 Explanatory Data Analysis

While starting the explanatory data analysis we soon realized that visualizing the complete Stack Overflow network is not possible. Therefore, we broke down the network into its different communities based on the tags in the edge list. To analyze those smaller networks we imported them into NetworkX[link]. To get an understanding of the networks we plotted common network metrics presented in Table 1. We stopped the calculation a layout for the network after a run time of more than 10 minutes, since it was obvious that we will not be able to use it for an interactive visualization.

Since this network was still to complex we reduced it even further. We reduced self loops of users as they have no impact on the community. Furthermore, we decided to visualize only the giant component. Therefore we redid our explanatory analysis with the reduced network.

As also the calculation for the layout for that network took too long and no meaningful insights could be observed we decided to change our approach. As already mentioned we then focused on the behavior of different users instead of their connectivity. To get a first understanding of the users of Stack Overflow we visualized the distribution of collected features via histograms and scatter plots.

We observed that the features are distributed by power law. Therefore we decided to classify users into different categories. As classification we applied a simple threshold classification based on reputation using the intervals from the histograms:

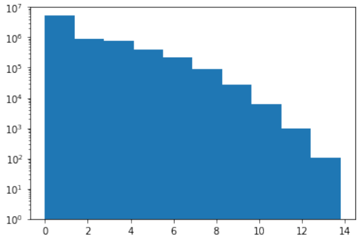
$$[1, 4), [4, 16), [16, 248), [248, 985), [3907, 15508), [15508, 244257), [244257, 969386)$$

To analyze the behavior in these user classes we calculated then the average features of users per class presented in Table 2.

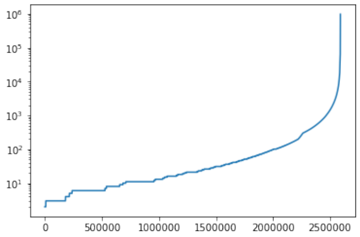
The values from Table 2 clearly show that users behave differently on Stack Overflow. Based on these behavior we will call their classes:

1. Inactive Users

2. One Time Users



(a) Distribution of user logarithmic scaled reputation



(b) Scatter of ordered user repuatation

Table 2: My caption

Feature/Users	Inactive	One Time	Active	Frequent	Super
Number of Users	5240899	1635527	615045	118480	7240
Reputation	1.09	19.63	258.19	3092.1	43666.3
Votes	-0.03	0.91	13.85	155.61	1542.17
Questions	0.05	0.97	4.93	21.1	47.03
Answers	0.06	0.58	4.98	42.05	444.21
Average Answer Votes	-0.01	0.24	1.19	2.04	2.79
Average Question Votes	-0.02	0.38	1.83	5.99	13.36

3. Active Users
4. Frequent Users
5. Super Users

In the next chapters we will explain various visualization techniques in order to visualize these differences.

5 Design

After deciding which questions we wanted to answer and who wanted to present the insights gained by analysis and visualization, we imagined how the audience would navigate our visualization.

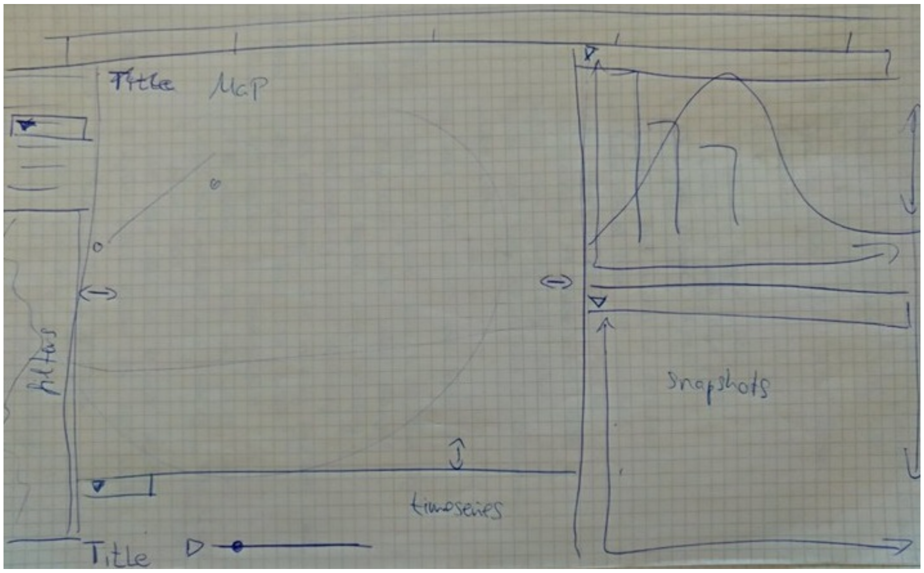


Figure 2: Dashboard Wireframe created for network visualization

We started out by sketching a dashboard like wire frame with an interactive network graph in the center panel, sidebar to the left including filter settings in drop down menus or open list selections and a side panel on the right side hosting 2 distribution plots that would changed based on the filters selected and clicks placed on the network in the center panel. In addition to that we wanted to demonstrate how different communities developed over time, by putting a time line with a play button in bottom of the that allowed the user to scroll to certain points in time or have it play over the complete time range so he would be able to see the network and the distributions changed over time.

As outlined in the section above we quickly realized that an interactive network plot of this size would not be feasible and even focusing on smaller communities centered around libraries like pandas or d3 would still be extremely

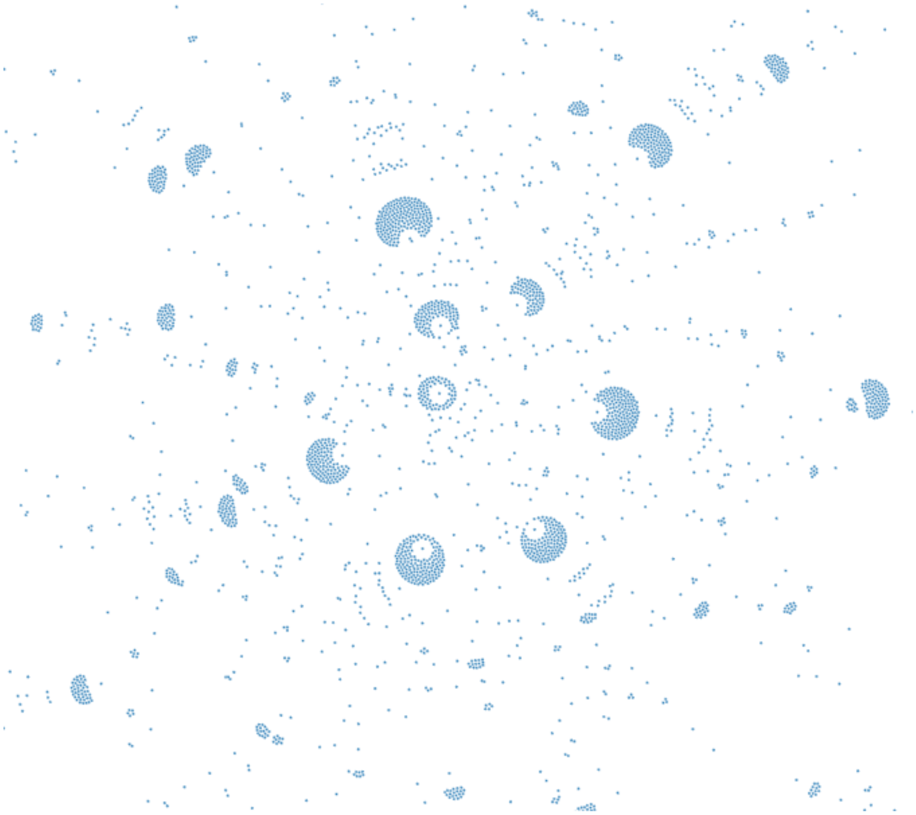


Figure 3: Network created by force directed layout for cmd community

challenging to realize in a smooth manner. The graphs, like the one above, would take a long time to render and the interactions would be clumsy and stuttery. Therefore the idea to make strongly engaged users stand out by highlighting their high number in connections in a graph network was abandoned.

After reconsidering how we could highlight these "Super Users" on the platform we decided to take a statistics based approach emphasizing more quantitative measures instead of a relational plot showing the position in a graph. By classifying the users into 5 classes we were able to bring back some features for qualitative differentiation without cluttering the plots to much. This was further supported by picking colour schemes that used a differences in hue to illustrate changes in magnitude or intensity and apply different colours in case of qualitative changes.

and design the project along more of a story line than single site dashboard.

How can one visualize these different users. Which things do we want to show: Difference between users, users as such, evolution of community, evolution of users

Explain chart choices Bar charts Describe why we picked - perception and visual encoding Keep it simple, show only one attribute Do not use Pie charts since it is hard to observe areas sort for readability

PA Describe why we picked - perception and visual encoding show several items with feature correlation Show first then evolution to final: Reorder, log scale

And which evolution over time can we observe

Streamgraph Why did we pick - perception and visual encoding Show evolution of quantative values over time Compare quest and answers directly in one chart Show first then evolution to final, normalize

Bubble Evolution Why did we pick - perception and visual encoding Evolution of items over time: Color (category to seperate), two other qualative attributes Show first then evolution to final, switch

CREATE sketch for step by step approach to guide user, page layout

Interactions: Later idea: Plot the user to involve him

Happy Testing with users

6 Implementation

The histograms for reputation and number of questions and answers asked show the discrepancies between different Stack Overflow users. It is possible to switch between features using the input box on the upper right [Figure FeatureSelector]. E.g. the histogram of the answers asked shows that there exist more than 6 Million Users how have never answered a question while there are a few hundred users who have answered more than thousand answers. This trend is easily observable without interactivity. However, to give the user more details we

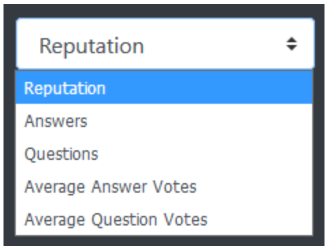


Figure 4: Select statistic to display by selecting a feature

have added a tool tip when hovering over bars[Figure]. The tool tip shows the number of users in the corresponding bin.

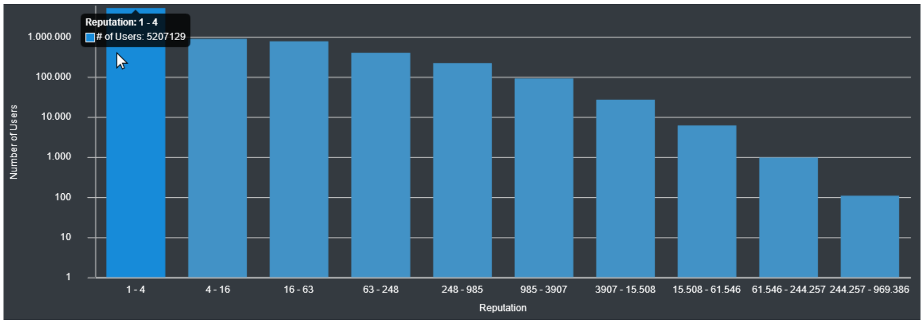


Figure 5: Tool tip when hovering over bins to show more details

The horizontal bar charts shown after the classification highlights the difference between the user types. One can observe that the average values of features of user types vary heavily. E.g the number of answers posted as inactive use in average is only 0.06 while super users have posted in average more than 400 questions. Once again, the user can select the features to display from the selector. Furthermore, more the exact values of the bars can be retrieved with a tool tip.

The parallel axes visualization shows a sub selection of Stack Overflow users by visualizing their activity represented by our features. The chart shows the different correlations between classification and features. On one hand, One can notice that the features on the right hand side of the chart are highly correlated. This means that users with a high reputation are likely to answer many. One the other hand there exists no strong correlation between questions and our classification. It is possible to highlight a selection of users by brushing on the axes [Figure PASelect]

Furthermore, one can reorder the axes in order to highlight correlations between classes.

In the following chart we are presenting the evolution of the Stack Overflow content. In the first years after the creation of Stack Overflow questions and answers were mostly posted by super users. But the chart shows two different

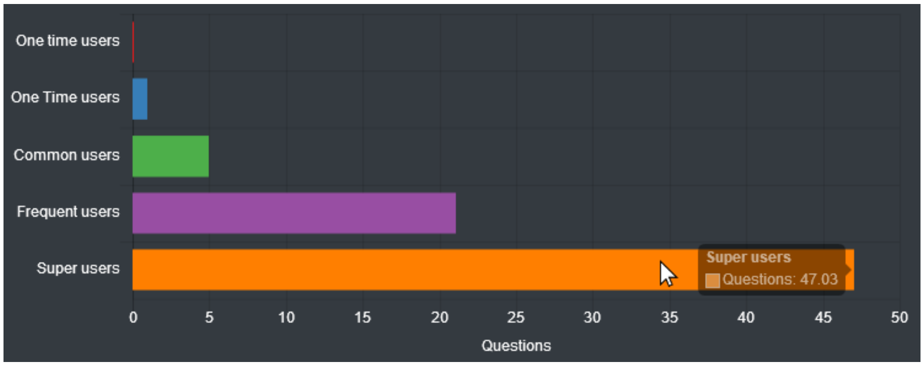


Figure 6: Tool tip when hovering over average statistics per class

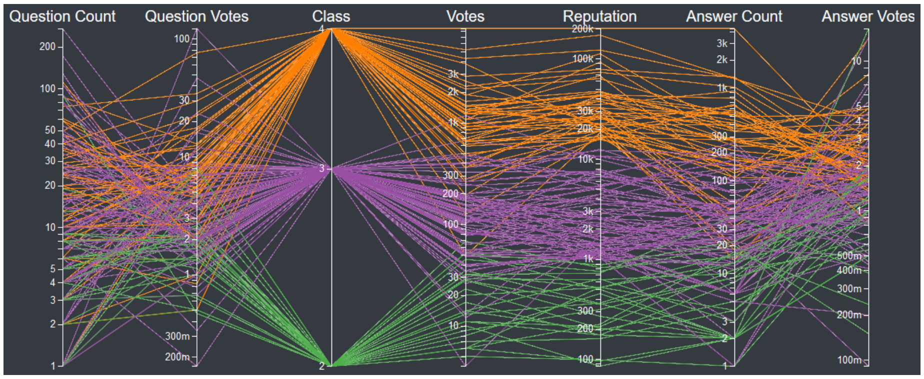


Figure 7: Select users via drag selection on the axes

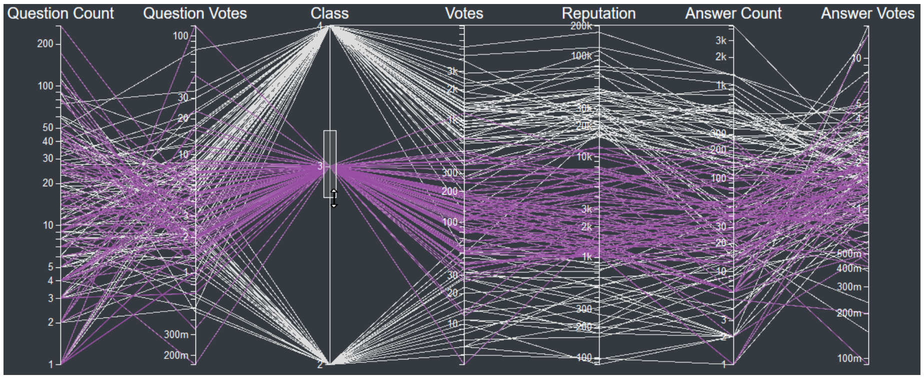


Figure 8: Select users via drag selection on the axes

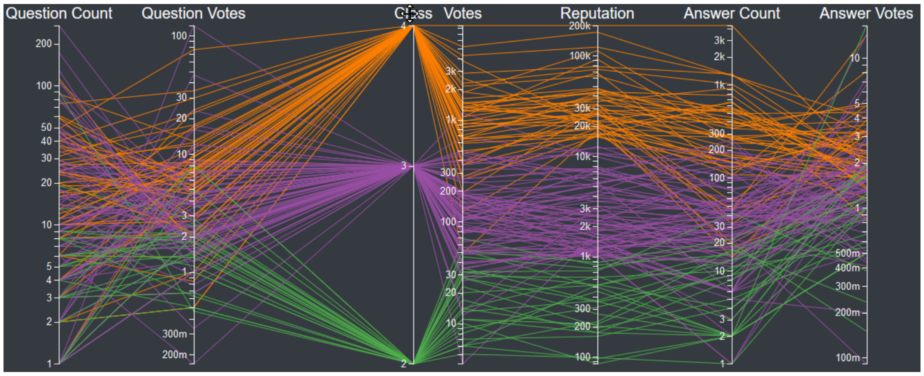


Figure 9: Reorder axes by draggin them to a new position

trends for the evolution of questions and answers. While super users are still answering almost the same portion of questions, questions are asked by users from other classes. Therefore, Stack Overflow nowadays relies on super users to answer and on less frequent users to pose questions. The exact evolution of portions is displayed in a tool tip when hovering over the chart area.

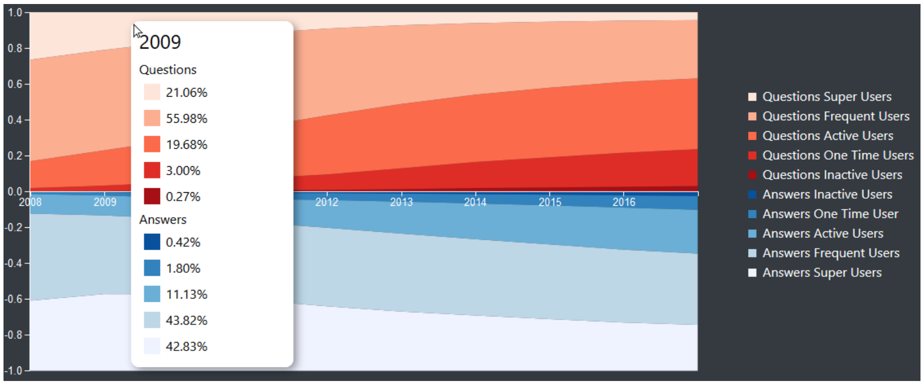


Figure 10: Tool tip when hovering over Stream Graph to show distribution of one year

In the last chart of our visualization we want to visualize the evolution of users. When pressing the play button or using the slider [Figure BYear] one can observe the change of user types over the last ten years. When observing the size of the bubbles it is observe able that the number of users in the less active classes increases much faster than for more active users. But since we pointed out that frequently active users are crucial for Stack Overflow let's have a look at their evolution: The number of questions asked is increasing from 7 to 47 by a factor of 7. During the same time the numbers of answers posted by factor of 60. The same trend is more likely to apply the more active a user is. Therefore one can say, that people get super users by answering more and more questions. The bubbles for the different user types can be hidden by clicking on

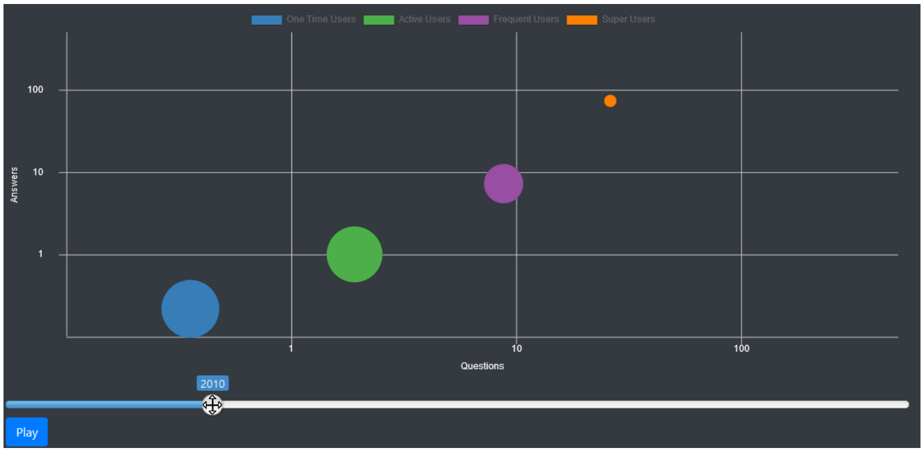


Figure 11: Change the year of the bubble chart by moving the slider or by clicking play

their corresponding label in the legend.

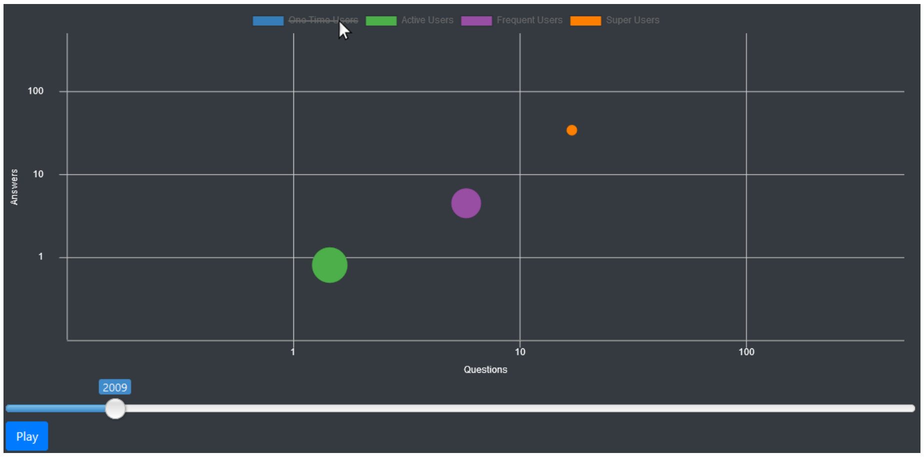


Figure 12: Hide classes by clicking their label on the legend

7 Evaluation

Show trend in bubble plot Search by User Name Create own classification Click on bubble in plot shows subselection of users

8 Peer Assessment