Data Visualization
Final Project Process Book

Class Brüß and Maximilian Rünz

December 2017

1 Introduction

Stack Overflow has grown to one of the most popular websites not only for computer scientists, but also for the public of coders with 50 Million unique monthly visitors. Since it has become a very important tool for developers we were curious about who actually helped to keep this community running. 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 improve their contribution.

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 grow 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 grew to become such a popular platform. Additionally 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 project mainly addresses Stack Overflow users and aims to provide them an insights about their role in the community by put them into context with the larger engaged community. In addition we would like to show how much work is put in by the users that really carry the community.[1]

[1] "Currently limited to selected users"

2 Related Work and Inspiration

The idea of analyzing Stack Overflow came up since during our daily life when 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 variety of users on the platform. Since we were also thinking about visualizing communication patters of humans we decided to analyze 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 illustrating the development and growth of the community we were mainly inspired by Gapminder[https://www.gapminder.org/tools/] showing a very expressive and interactive time evolution of three quantitative values and one categorical value.

3 Dataset and Preprocessing

The analysis and visualizations are based on a data dump from Stack Exchange posted 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 global stackoverflow.com. These 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, 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

Table 1: Network metrics JavaScript community

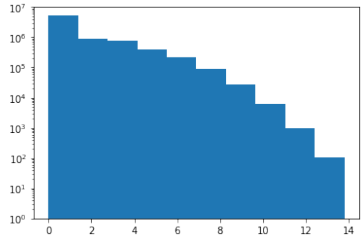
Feature	# Nodes
# Nodes	417219
# Edges	765669
# Conn. Components	36140
Avg. Node Degree	3.67
Avg. Clustering	0.0014
Giant Component	357624

of questions and answers posted, the average votes for questions and answers and the number of votes of all questions and answers posted.

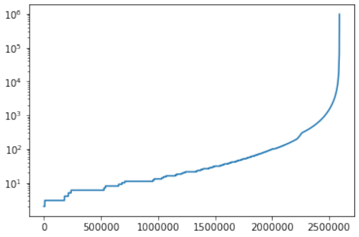
4 Explanatory Data Analysis

While starting the explanatory data analysis we soon realized that building an interactive browser based visualization on 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. To get an understanding of the networks we plotted common network metrics presented in Table 1. We stopped the calculation of 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 removed 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.



(a) Distribution of user logarithmic scaled reputation



(b) Scatter plot of ordered user reputation

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

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

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

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

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
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 to which audience we wanted to cater, we imagined how this audience would navigate our visualization.

We started out by sketching a dashboard like wire frame imagining an interactive network graph in the center panel, a sidebar to the left including filter settings in drop down menus or open list selections. Additionally a side panel on the right side was intended for hosting two distribution plots that would have 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

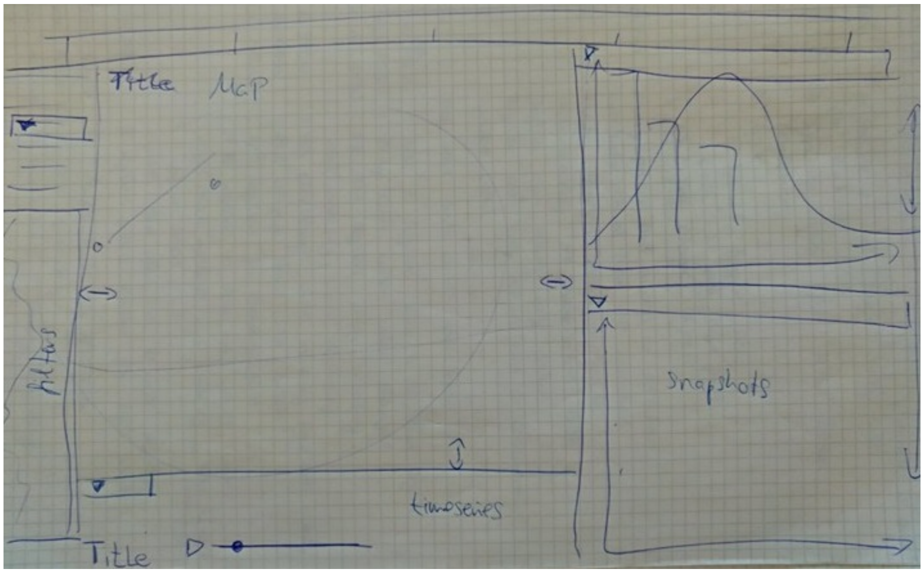


Figure 2: Dashboard Wireframe created for network visualization

communities developed over time, by putting a time line with a play button in bottom. This would have 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 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.

We also decided to move from a single site to multiple sliding pages since not all plots and graphs are connected through interaction and in order to reduce information density to a more comfortable level. We used to shift to inform the user in a more story telling oriented way instead of a plain all in one page. To

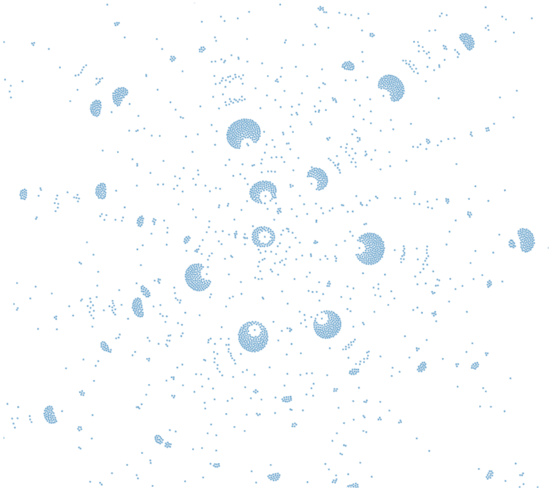


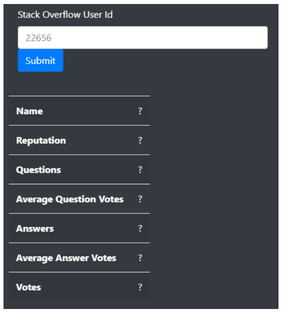
Figure 3: Network created by force directed layout for cmd community

enhance the experience even more we decided to the take user along by allowing to enter the user's Stack Overflow User Id and see himself being placed in the midst of the rest of the community.

5.1 Motivating Chart Styles and Choices

In the following subsection we will motivate our final choices page by page and discuss which option were replaced.

5.1.1 Have you been active on Stack Overflow?



Stack Overflow User Id

22656

Name	?
Reputation	?
Questions	?
Average Question Votes	?
Answers	?
Average Answer Votes	?
Votes	?

Figure 4: Streamgraph showing percentage of contribution per user class over time

With this very reduced page we intend to just quick receive the Stack Overflow User Id from our user, in case he has one, show some very basic stats on his profile. Entering an id is not mandatory.

5.1.2 User Contribution

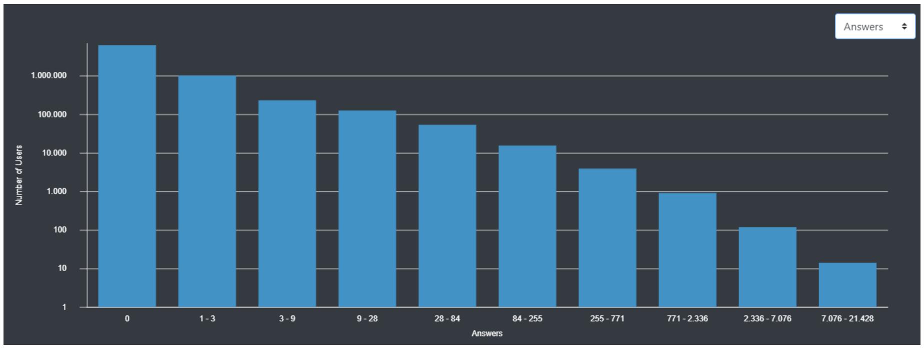


Figure 5: Streamgraph showing percentage of contribution per user class over time

This page includes one panel which can be set to show one of three histogram plots, show the distribution and range of the metrics for contribution (reputation, number of answers posted, number of questions posted). We choose these very clean plot to set up the plots on the following pages by providing a basic overview of the distribution of the mentioned metrics over the user base as a whole.

5.1.3 User Classification

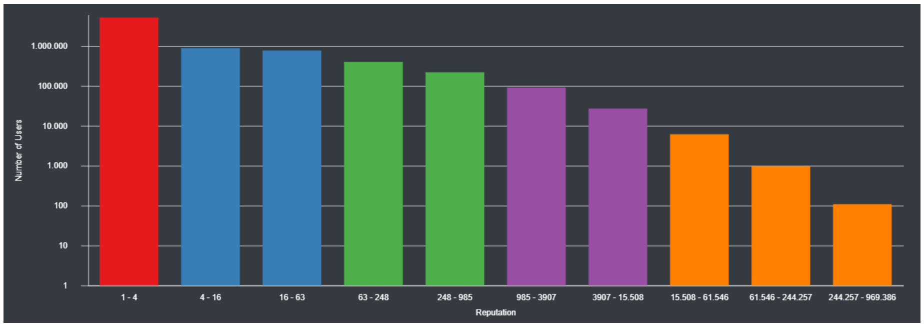
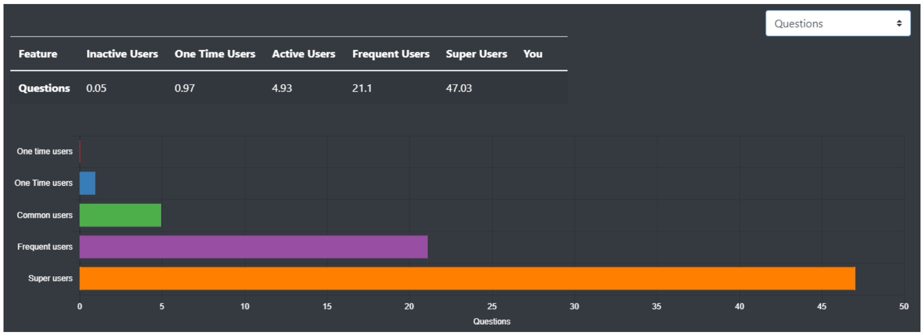


Figure 6: Streamgraph showing percentage of contribution per user class over time

This is the first page where we show more than raw data, by classify users into 5 classes shown in distinctly different colours. It is imperative that this separation is made clear since every following plot or graph, except for the stream graph, reuses this exact colour way to allow our audience to follow long without being confused.

5.1.4 User Behavior in Classes



Feature	Inactive Users	One Time Users	Active Users	Frequent Users	Super Users	You
Questions	0.05	0.97	4.93	21.1	47.03	

Figure 7: Streamgraph showing percentage of contribution per user class over time

Now that user classes are established we would like to characterize them. We do this by comparing the user classes in 5 different metrics, such as answers posted. The horizontal bar chart allows the audience to get a quick intuitive impression about how much the user class averages differ from one another in certain metrics. Since the Averages are used in last plot we wanted to clearly state them. This also allows the user to gain inside into whether is above or below average in his user class. Initially a doughnut shaped pie chart was used in place of the bar charts, but was ultimately removed because the differences between the user classes being too unclear.

5.1.5 Individual Users

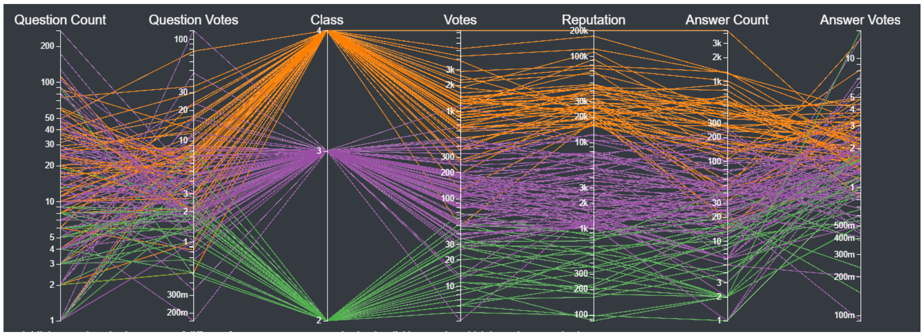


Figure 8: Streamgraph showing percentage of contribution per user class over time

This parallel axis chart represents single users from the three most engaged user classes by single lines keeping the set colour way for the user classes and highlighting the user with the entered User Id in yellow. In the inital order the axis 'Class' represents the center. To the right of the center we positioned the axis for the metrics that show a rather clear correlation to one another, while to the left the axis for the metrics with no clear correlation were placed. More interactivity is described in the section 'Implementation'.

5.1.6 Class Contribution over Time

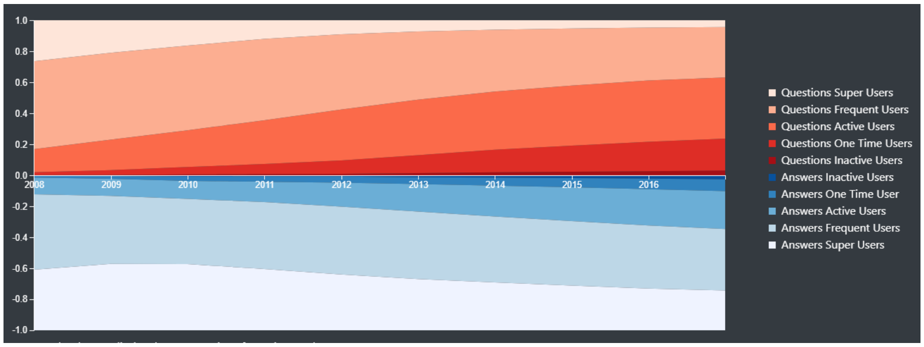


Figure 9: Streamgraph showing percentage of contribution per user class over time

Since we wanted to illustrate the changes and growth of the community, which we attempted initially through a network graph plot changing over time, we decided to split this into two charts two pages. The first chart is a stream graph showing the development of the share in question and answers posted by each user class over the years. The double sided layout(red on top, blue below) to allow the direct comparison between the change in questions and answers posted.

5.1.7 User Evolution

This second chart is a bubble chart inspired by gapminder.org . Like Hans Rosling in his famous TED talks, we wanted to create an animated graph showing the evolution of the user classes through the years. The bubbles are again kept in the colour way that was introduced on the 'User Classification' page.
The bubble size shows the number of users in the respective user classes. We chose to represent this metric, because we were less interested in communicating the exact pace of growth and were looking to emphasize that is platform has grown immensely and that the less engaged user classes have grown more quickly. This graph was tuned several times to make this more obvious. In addition we choose to illustrate the user with the entered Stack Overflow User Id separately, since we considered this look back a good way to close out our narrative.

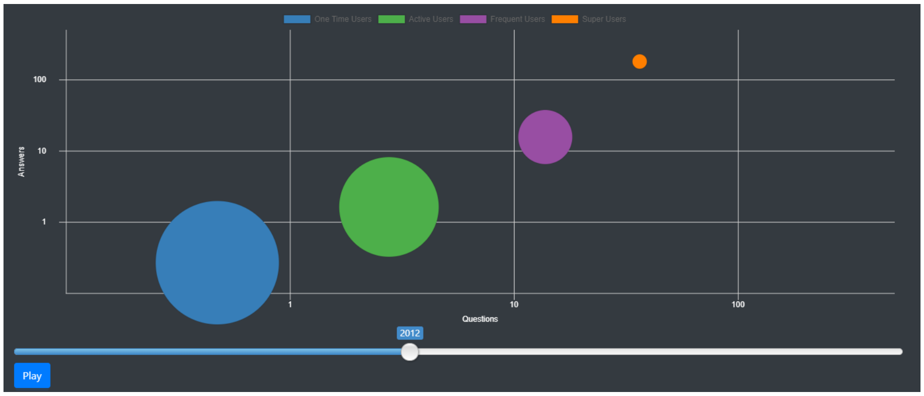


Figure 10: Bubble Chart showing growth and shift of question to answer ration over time

6 Implementation

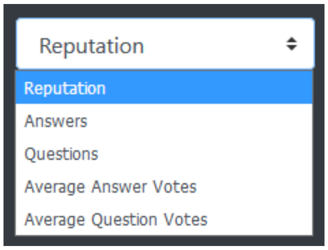


Figure 11: Select statistic to display by selecting a feature

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 Feature-Selector]. 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 on demand we have added a tool tip when hovering over bars[Figure]. The tool tip shows the number of users in the corresponding bin.

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, the exact values of the bars can be retrieved with a

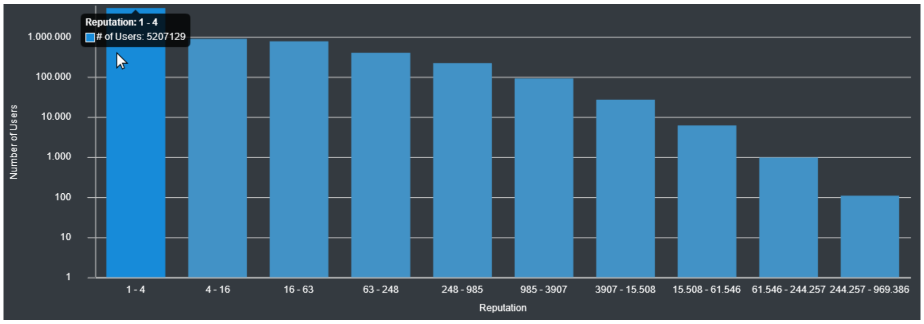


Figure 12: Tool tip when hovering over bins to show more details

tool tip.

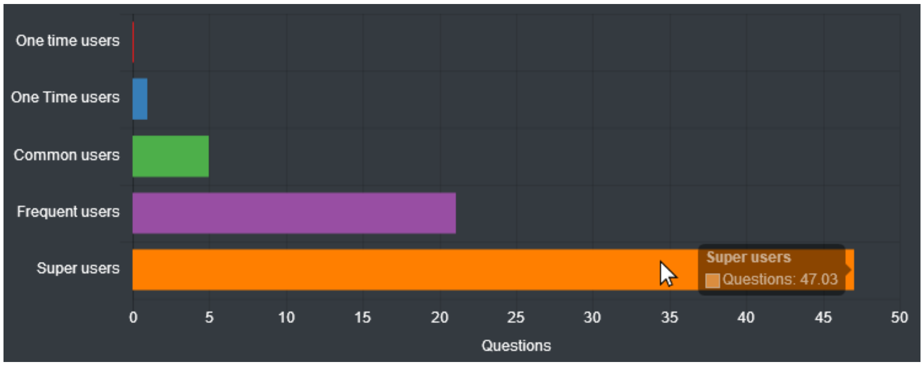


Figure 13: Tool tip when hovering over average statistics per class

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 post many answers. On 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 other correlations between features.

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 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.

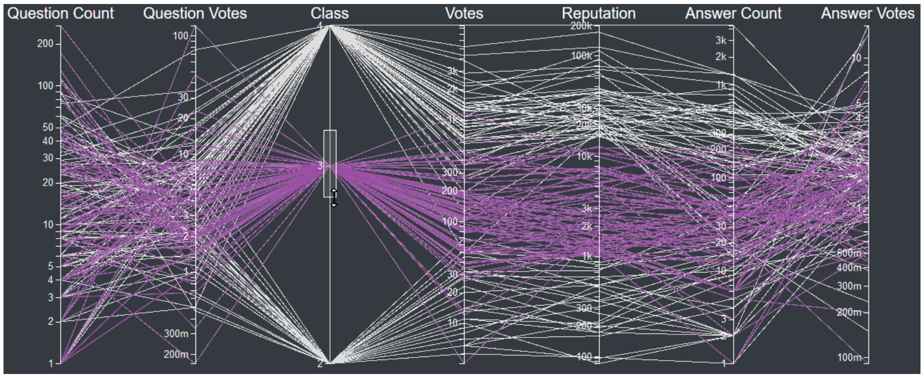


Figure 14: Select users via drag selection on the axes

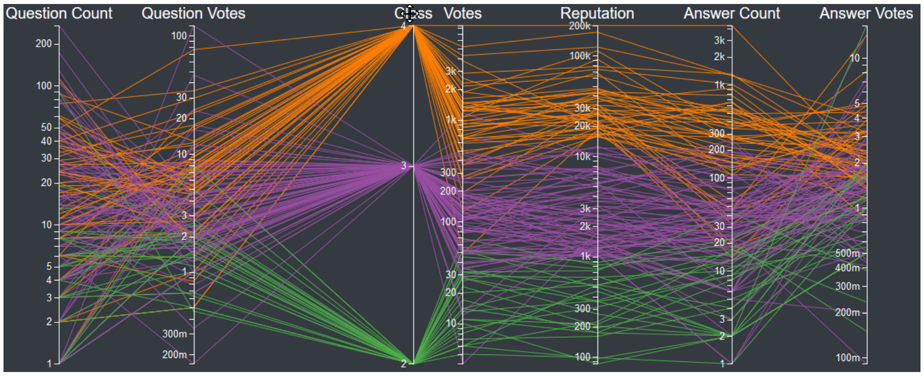


Figure 15: Reorder axes by dragging them to a new position

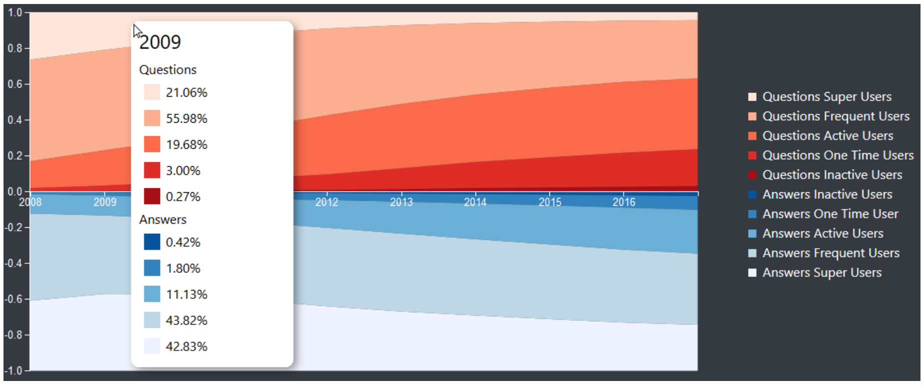


Figure 16: Tool tip when hovering over Stream Graph to show distribution for selected year

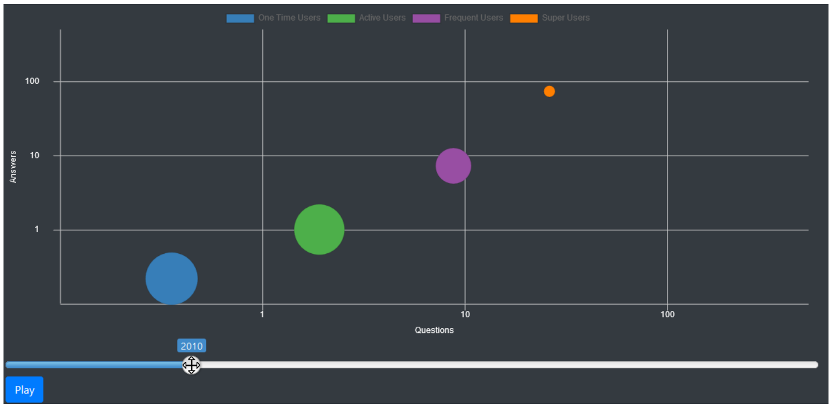


Figure 17: Change the year of the bubble chart by moving the slider or by clicking play

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 viewable 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 their corresponding label in the legend.

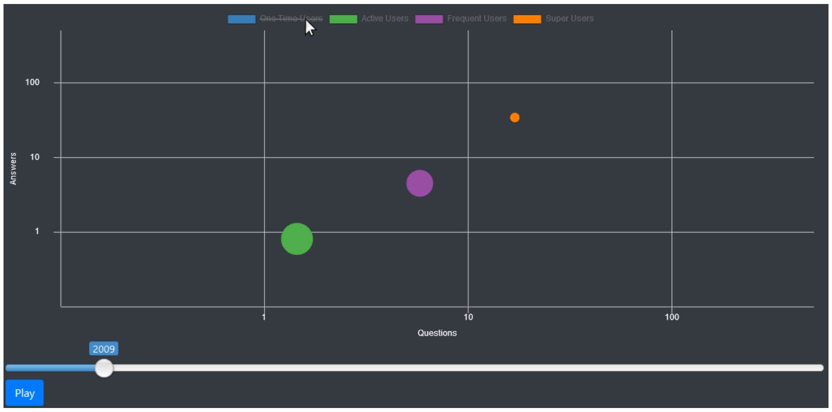


Figure 18: Hide classes by clicking their label on the legend

7 Evaluation

We have shown that there exist different types of users: inactive, active and very active users. Questions on Stack Overflow are nowadays asked by every active user. Answers, on the other side, are posted by very active users. They contribute 65% of the answers, while being a minority of 2%!

Furthermore, the visualizations showed that Stack Overflow content shifted over time. Super Users were asking many more questions than today. Nowadays they spend their time answering questions asked by less frequent users.

Last but not least we were able to observe the growth of super users: Super Users shifted from asking questions to answering hundreds.

Remembering that we did not spend a single word on non-registered users, makes the contribution of active users even more impressive. According to Quantcast, Stack Overflow has almost 50 Million unique visitors per month!

8 Peer Assessment

8.1 Peer Assessment for Claas Brüss

Preparation – were they prepared during team meetings?

The meetings were productive, new approaches assesed and discussed. Several times agreed deadlines were not met, it took several more days until results were available.

Contribution - did they contribute productively to the team discussion and work?

New visualisation approaches were discussed. However the contribution to the code was missing: No contribution to the code of the website. Small contribution to the data processing and analysis

Respect for others' ideas - did they encourage others to contribute their ideas?

Yes

Flexibility - were they flexible when disagreements occurred?

Yes

8.2 Peer Assessment for Max Rünz

Preparation - were they prepared during team meetings?

Yes. If couldn't find a creative solution he was sure about he came prepared with ideas where to look further. Team meetings always yielded a list of steps to be taken next.

Contribution - did they contribute productively to the team discussion and work?

Max contributed the bulk of the code and therefore compensated for me being far less experienced than him. His input in the discussions prevented us from rushing into technically unfeasible ideas multiple times.

Respect for others' ideas - did they encourage others to contribute their ideas?

The discussions were always fairly open, enjoyable and rarely left in disagreement.

Flexibility - were they flexible when disagreements occurred?

Usually yes. Only in really high stress situation during the last weeks with the deadline crunch before Christmas approach we had some disagreements about how priorities should be put, but those were in no way too severe.