Stack exchange network analysis

Miroslav Král
MSc Business Analytics
University College London
Miroslav.kral.15@ucl.ac.uk

Agus Nur Hidayat
MSc Business Analytics
University College London
Agus.hidayat.15@ucl.ac.uk

1 Introduction

The purpose of this paper is to summarize the result of our analysis of social networks in an online Q&A community **stackexchange.com**.

2 Dataset description

In this paper we work with online community data, specifically with data from sports stack exchange community. Stack exchange is a network of question answers web communities on various topics. One of the pages (communities) that belongs to stack exchange portfolio is page on sports.

Sports stack exchange is question - answer web page, that means that users can specify questions and other members of community can answer these questions. There can be multiple answers to one question, however, user, who specify question can then mark answer that satisfied his information need as accepted one. Users are allowed not only writing a posts, but comments too. Comments can be related either to question and answer. Stack exchange also implemented voting system that enables to all users rate questions and answers (both together are referred as a posts) by plus and negative points. These points are then transformed into reputation of user who wrote the post and add a gamification element to the page. Great benefit of Stack exchange pages is that they are opened to everyone, not only for registered user.

We acquired a dataset that contains all the data about activity of the users of Sport stack exchange web page. This dataset is released periodically by Stack exchange through following link: https://archive.org/details/stackexchange. It contains all the data since beginning of this community (February 2012) until March 2016 (date of the last release).

Sports stack exchange community belongs to the smaller communities Stack exchange portfolio; therefore, the dataset is reasonable big. Community consists of 5018 users whose wrote a 2676 questions and 4353 answers. Data set consists of 5 tables. The most important tables are table Users and table Posts.

Table Users capture information about users of stack exchange web site. All the user related data, such as name, location of user, when he joins community, when he last access page and other more. However, the most important feature of user is reputation that capture how are the user's contributions accepted by community.
Table Post capture all the post that users wrote. It contains text of the post, who wrote it, when wrote it, how many positive and negative voting post received and most importantly it contains feature whether it is a question or answer. In case it is a question it is also captured accepted answer (if exists) and in case of answer is denoted to which question belongs. In the figure below we can see how number of post per month depends on time.

Figure 1 Number of posts per month

X-axis on the plot above capture time and y-axis shows number of posts. Generally, we can see that on average wrote around 150 posts (question and answers together) per month. However, there are high peaks in February 2012 (almost 450 posts) and July 2014 (in my opinion this is caused by Football championship). The significant drop down in the March 2016 is caused by fact that data set contains record only for first few days of March.

Other tables are called Comments, Tags and Votes. Detailed description of all tables and columns of the tables are presented in Appendix A.

3 Question – Answer network

In this chapter of a report we will examine network or users of Sports Stack exchange community. This network is based on relationship "who answer whose question". In general, we decided to divide this part into two subchapters. In the first subchapter we tackle with network that contains all answers for a questions and in the second part we look specifically on the network where only question and accepted answers are captured. The main aim of this chapter is to describe relations of the users in these networks, identify the influential users in network and compare these set of users we identify with ranking of these users.

3.1 Question – all answers network

In this part of the report I will describe network that represents relationship of users who wrote question and users who answer this question, all the answers for this question are captured. Overall, this network consists of 1956 users. Out of this 1956, 1206 of users ask at least one question and 1160 users wrote at least an answer. These numbers are really interesting in the light of total number of users registered on Sports Stack exchange web page - only 40 per cent of users actively

participate on the web page. We visualize the network in order to get initial idea of the network:

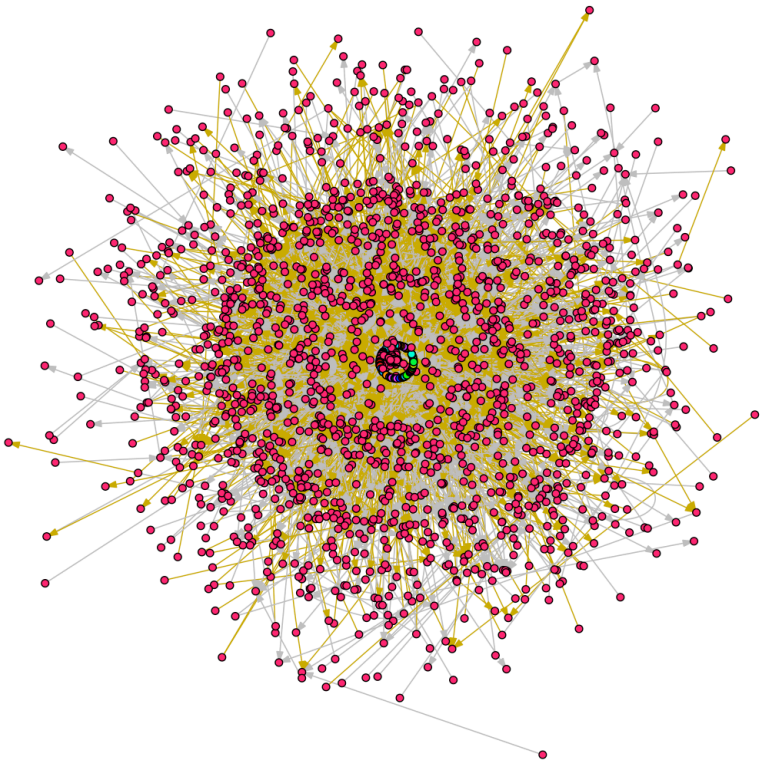

Figure 2 Network of all questions and answers

We can clearly see in Figure 2 that it is a big network with complex relationships and that due to the complexity it is difficult to get many insights from this Figure. However, there is still few very interesting insights:

1. Whole network is disconnected directed graphs. Colors of the vertexes detect connected subgraphs in the network. We can clearly see that most of the vertexes construct one big central connected subgraph, while there also exists few small subgraphs consist of small number of vertexes (for instance green nodes in the center of figure)
2. There exists two type of edges. First type of edges colored by gold color captures relationship of user who ask question and user who answer the question, but the answer was not accepted, while the second type of edge (grey) captures relationship of users who asked questions and users who wrote accepted answer to the question. It is not clearly visible from the network plot, but the is 4303 edges and only 1605 of them are accepted answers (network of only accepted answer and questions is presented further).

In order to get more insights of the network we plot the out-degree and in-degree distribution of this network.

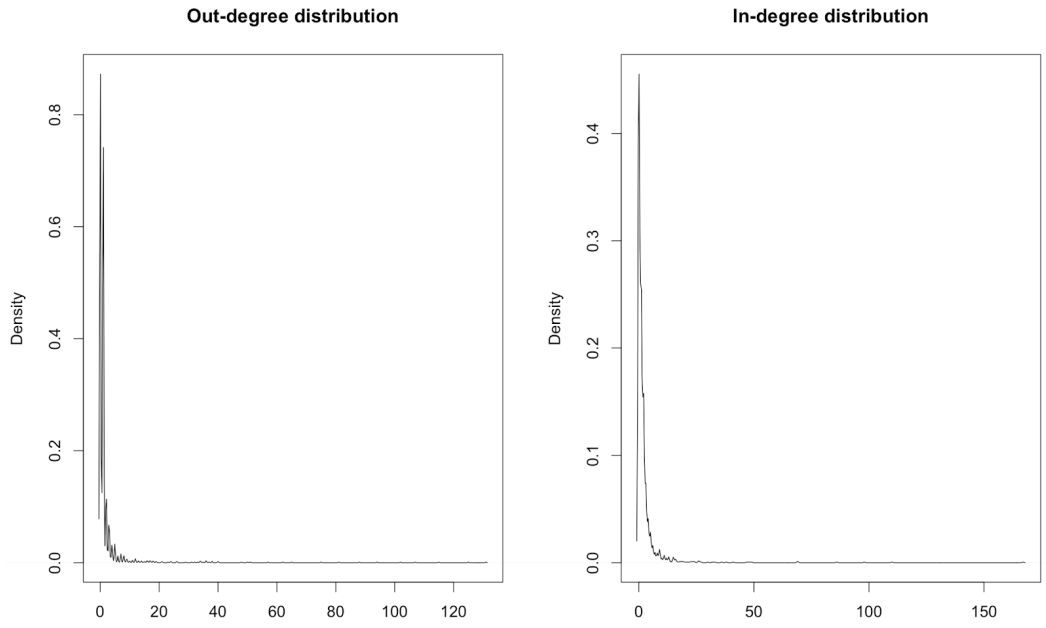

Figure 3 Out-degree and in-degree distribution

Figure 3 captures out-degree and in-degree distribution of a Q&A network. We can see that it is a fat tail distribution. Most users have out and in degrees small, between 0 and 5. These are the users who ask / answer just a few question and do not participate regularly in community. However, there are also users who have high in-degree and out-degree - the users who construct the core of a community and keep the community alive by actively writing answers / questions. We can see that maximal out-degree is around 120 answers and highest in-degree is around 130.

In the Figure 4 we saw that there are users who have have high out degree and in degree. In the plot below we looked who are these users.

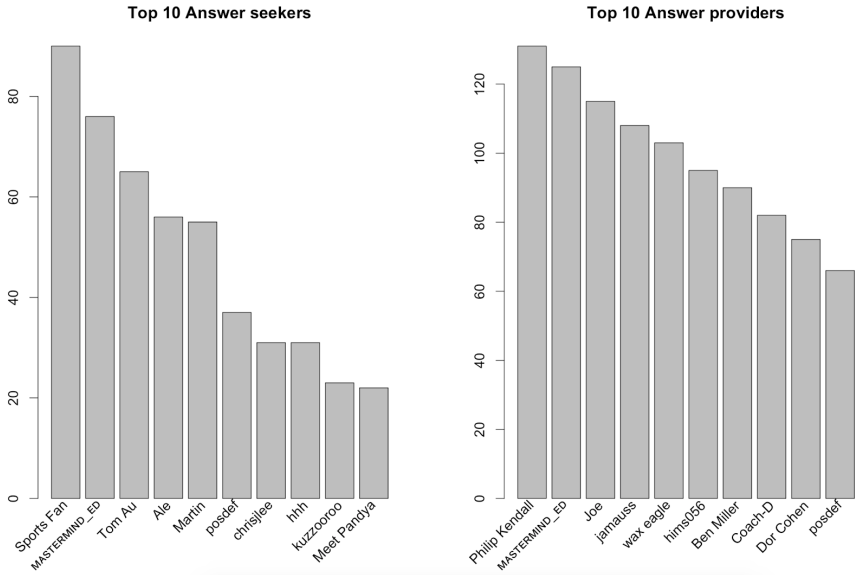

Figure 4 Most frequent answer seekers and providers

In the Figure 4 we can see who are the users who wrote most questions and most answers. We can see that only one name appears on both is MASTERMIND_ED who is also the user with highest reputation. Reputation of a user is a measure how important is user it a network. User can gain reputation by votes from other users. In the Figure 5 below we can see who are the users with highest reputation.

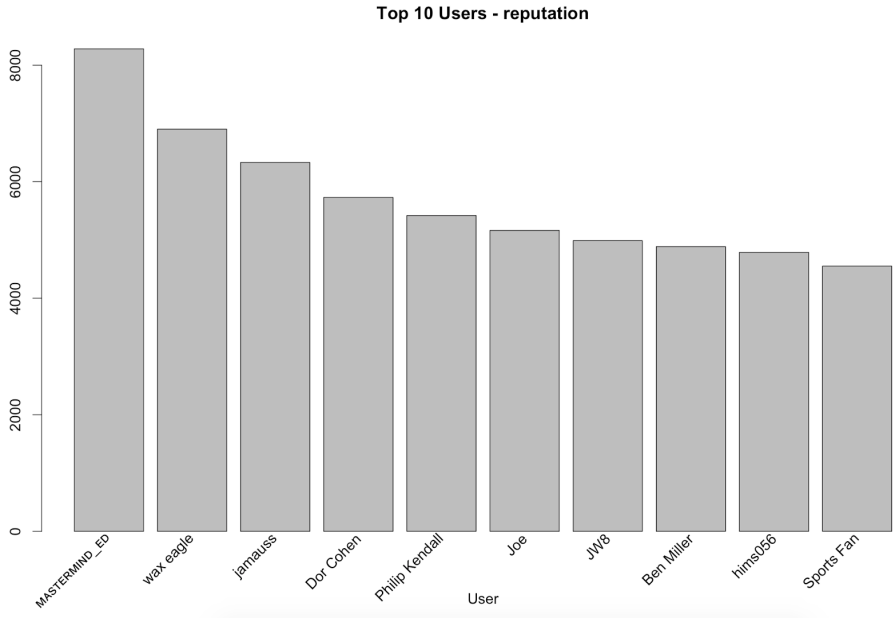

Figure 5 User reputation - top 10 users

We can see that 8 out of the most frequent answer providers are also listed as the users with highest reputation and only 2 of top answer seekers are in this list.

3.2 Question – accepted answer network

Question - answer network of accepted answers is the subset network of question - answer all network. Question - answer accepted contains relationship of users who wrote questions and users who wrote accepted answers for these questions. There is in total 346 users who wrote at least one accepted answer. This is really small number compared to total number of users which is more than 5000. In the questions - accepted answer network are also 664 users who ask at least one questions. These two numbers, number of users who wrote accepted answer and who asked questions is not distinct, a user can either ask question and answer questions (different one or either the same one). In order to get intuition about the network we visualize it. We can see the plot below.

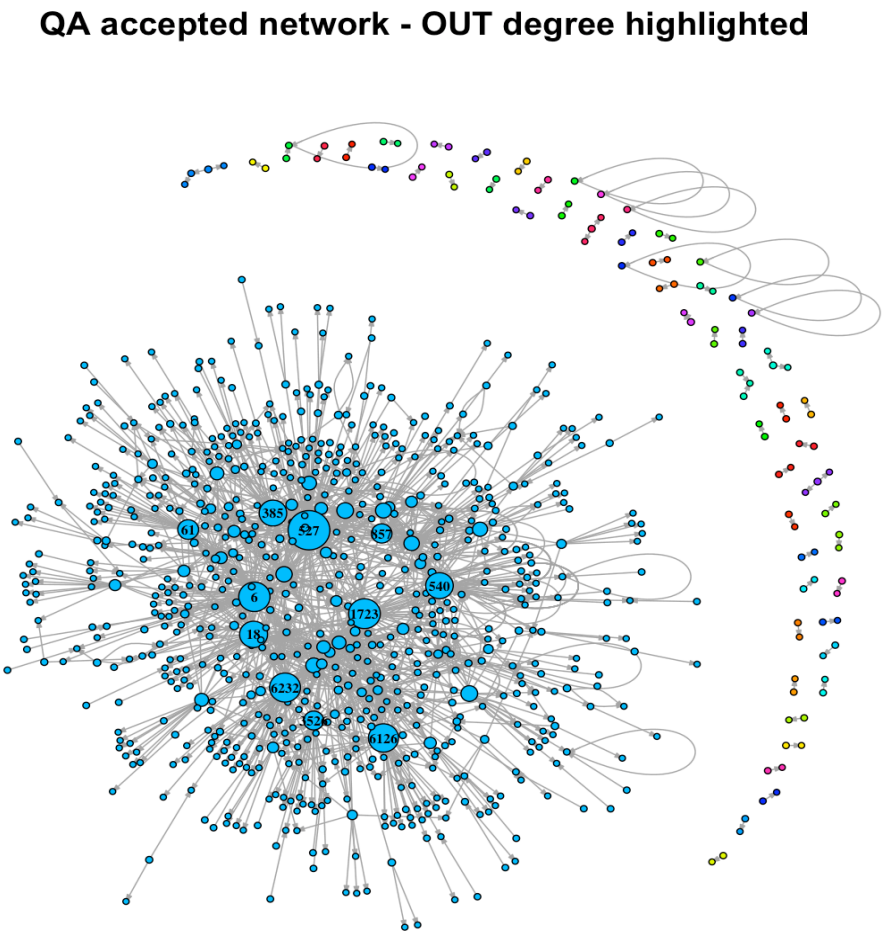

Figure 6 Q&A accepted network - OUT degree highlighted

The plot above capture question - accepted answer network. In this plot vertices represent users and edges of the directed plot shows who answer whose question. First thing we notice is that it is quite big directed plot. In the central part of the plot is large number of nodes denoted by blue color around this central group are large number of disconnected subgraphs. We can also see that the vertices have different colors, different sizes and some are labeled by number. Let's look what each of this features means.

1. Colors denotes communities in networks. We can see that the largest groups of nodes in center of plot marked by blue color is the largest community of the network. This community encompasses almost all the users. Rest of the network is formed by small communities with 2 or 3 members. We can also see were interesting insights, that there are communities with just one member. This means that there are users who wrote the question and who also answer this question and mark answer as accepted. That's
2. Size of the node is the most important thing captured in the network. It captures number of out-degrees that the node has, bigger vertex means higher out degree. In the other words we can say that vertex size shows how many accepted answers user wrote. This is in our opinion very interesting statistics because we expect that users who wrote a lot of accepted answers are the influential members of network. We can see very interesting, however not surprising insight. There is

large number of users who have answered only small number of questions and on the other hand there are a few users who has answer large number of question. These users should be the most influential ones in network (analysis of it further).

3. In order to identify who are the users with high number of accepted answers, we label the big vertices by numbers. These numbers are IDs of the users that unambiguously identify a user.

In the Figure 6 we identify users with high number of accepted answers. We can now verify that our hypothesis that high number of accepted answers means that user is important in the network. Note that user importance is measured by Reputation of a user.

In the bar plot below we can see who are the users with most answered and accepted questions from question - accepted answer network and how many question they answer.

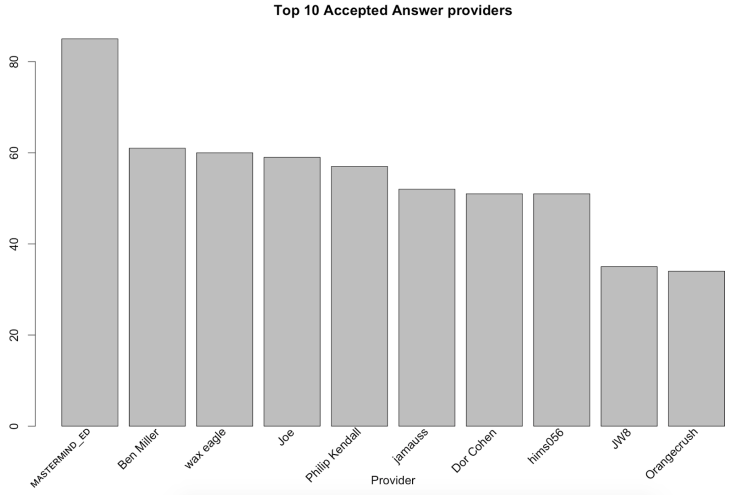

Figure 7 Top users with most accepted answers

From the Figure 7 we can see that user who wrote most accepted answers is called mastermind_ED and he correctly answer 85 questions. That's by far the highest number, rest of the top ten users answer 34 to 61 questions. We can also see that the users mostly overlap with the most active users overall presented in Figure 4. We can now compare this plot with the bar plot of users with highest reputation that is presented in Figure 5.

In the Figure 5 we can see who are the users with highest reputation. If we compare users in Figure 7 and Figure 5, we realize that most of the user names are the same and user called MASTERMIND_ED is either user with highest number of accepted answers and highest reputation. Overall, 9 out of 10 names are similar (however, the order is not the same). 9 out of 10 is higher number and it is evident that list of users who wrote most accepted answers overlaps better with list of top reputation users than the list of users who wrote most answers. We can also see that only the last member of top 10 accepted answer providers is not in top 10 users by reputation and vise versa. We can therefore conclude that our hypothesis that number of accepted answers means that user is important in network is correct.

Another thing we looked on was who are the users in this network who ask most of the questions, whether the set of users who have high number of accepted answers overlap with users with large number of questions and whether number of written questions that were answered is connected with reputation. First thing did was to visualize the network again. The network is visualized in a same way as the network in the Figure 6 (it is the same network in fact), however the difference is that we now highlight (by size of vertex) the users with highest in degree, that means users with high number of questions.

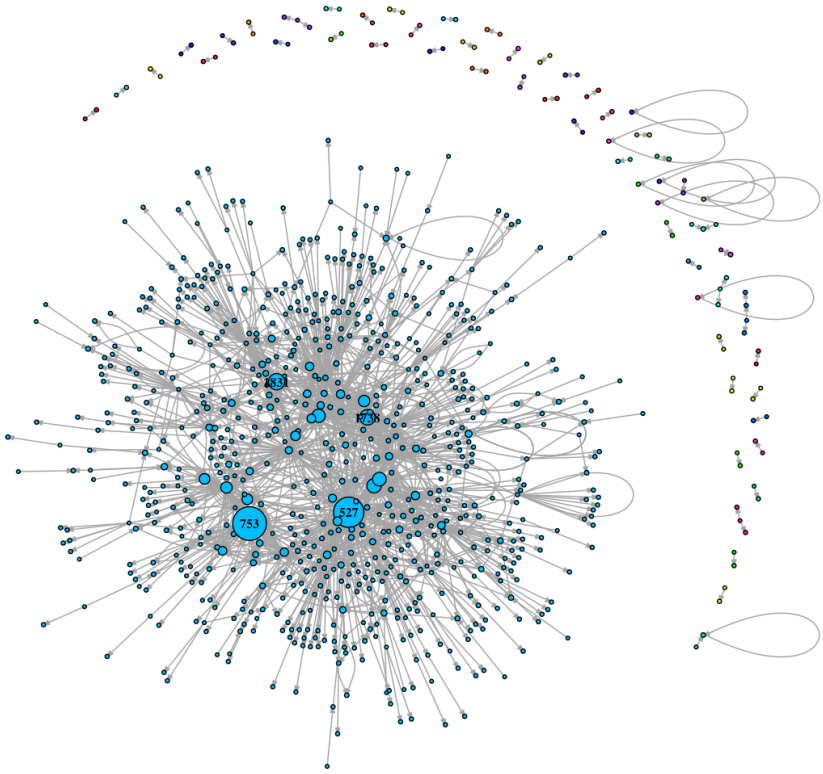

Figure 8 Q&A accepted network - IN degree highlighted

We can see that most all the characteristics of the network are still the same, high number of users asked small number of questions and only a few users ask questions frequently.

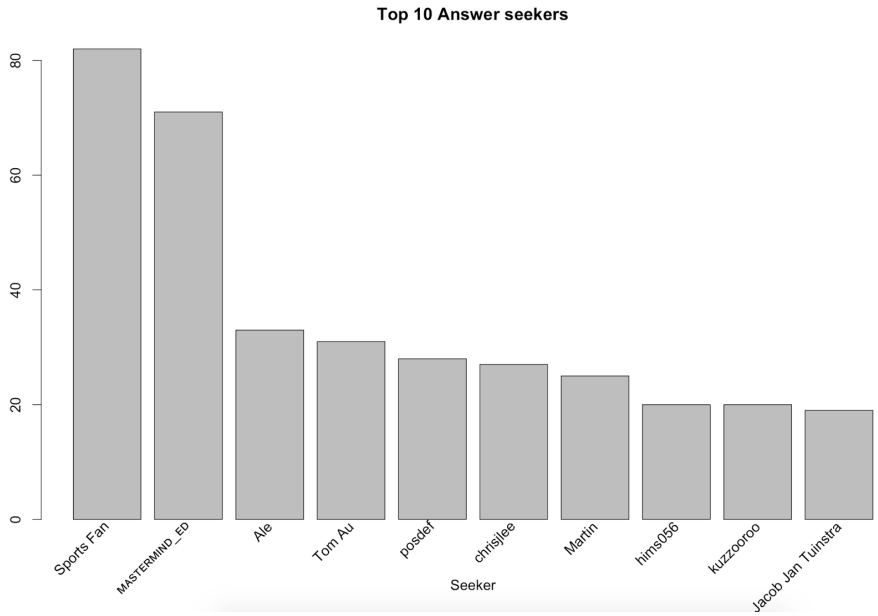

Figure 9 Top 10 answer seekers (question has accepted answer)

By comparing Figure 5 and Figure 10 we can see that only 3 users that are in top 10 answer seekers (with accepted questions) are also in the list of users with highest reputation. This shows expected insight that answering question correctly is better for gaining reputation than asking questions.
Overall we can conclude that we confirm that users who wrote most accepted answers are ranked as the top users of the Sports Stack exchange community.

4 Question – Answer network in time

Below are the networks visualization for Questions' Owners and Accepted Answers' Owners from year 2012 until year 2016. Basically network visualization in year 2016 is the same with network visualization for complete dataset. It happens because the network on each year is accumulated with the data from previous years. The first part illustrates in-degree of the networks which represents number of answered questions that each user has:

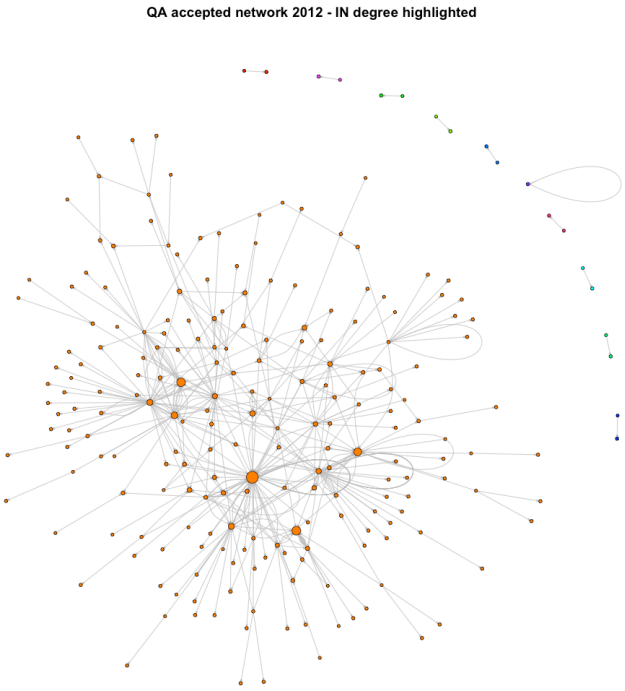

Figure 10 Network in 2012 (In-Degree)

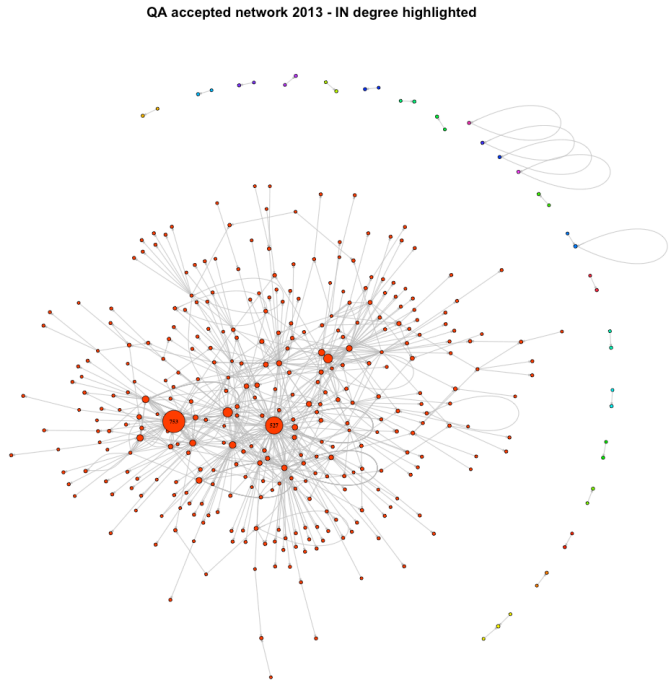

Figure 11 Network in 2013 (In-Degree)

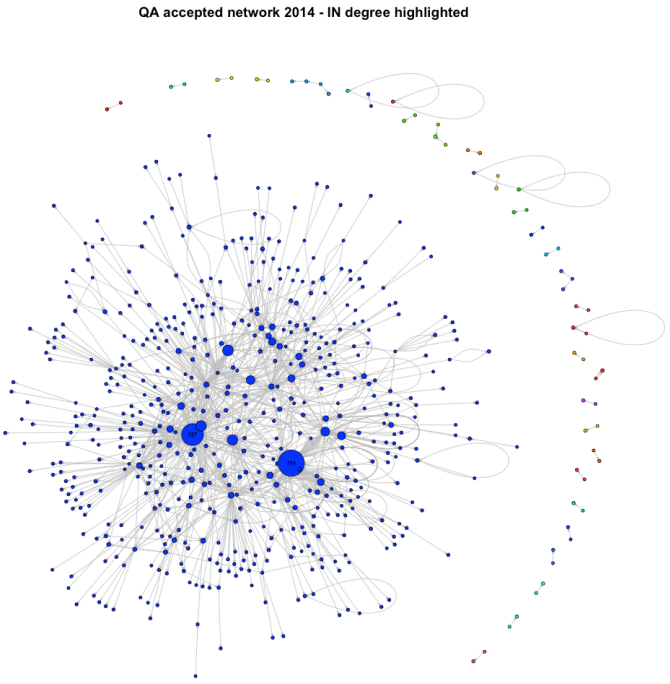

Figure 12 Network in 2014 (In-Degree)

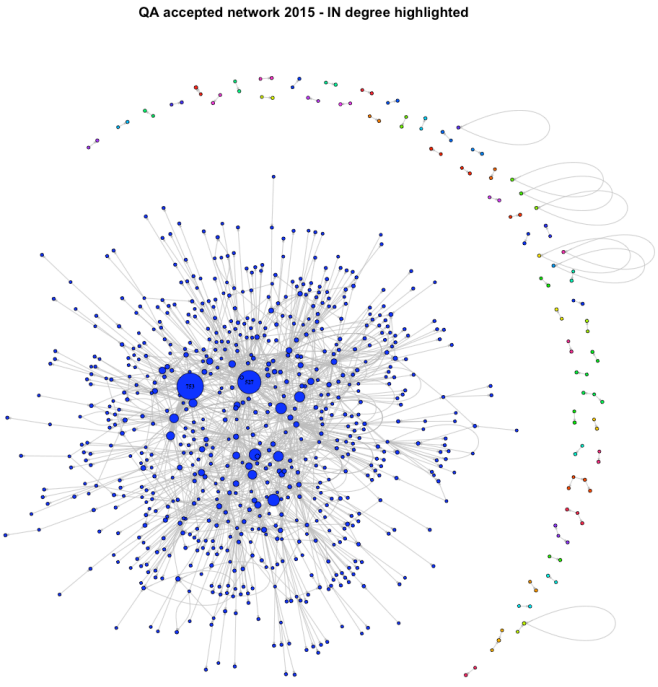

Figure 13 Network in 2015 (In-Degree)

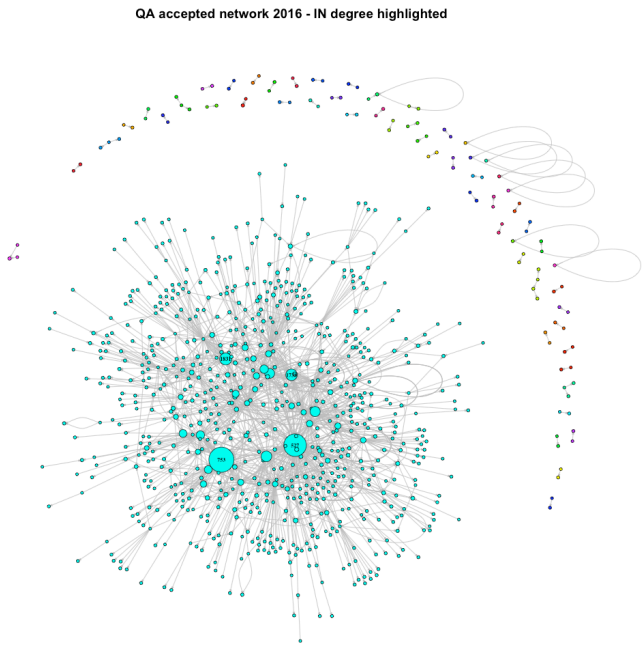

Figure 14 Network in 2016 (In-Degree)

The second part below illustrates out-degree of the networks which represent number of accepted answers that each user has:

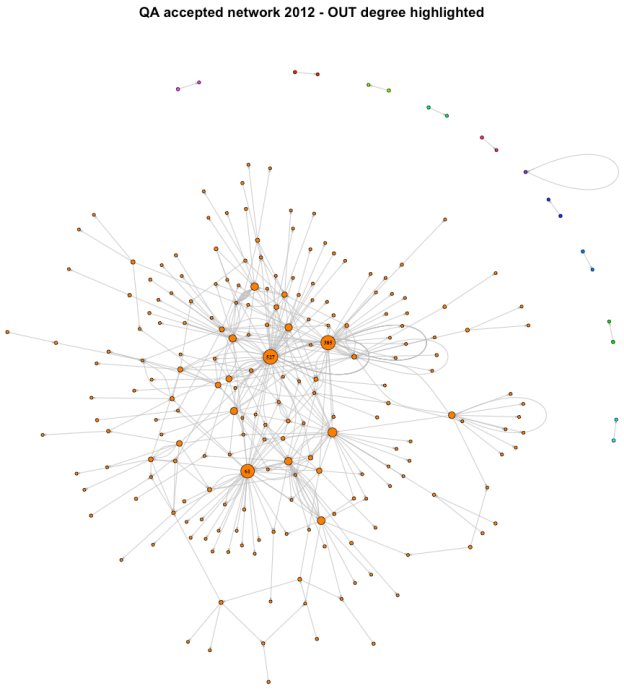

Figure 15 Network in 2012 (Out-Degree)

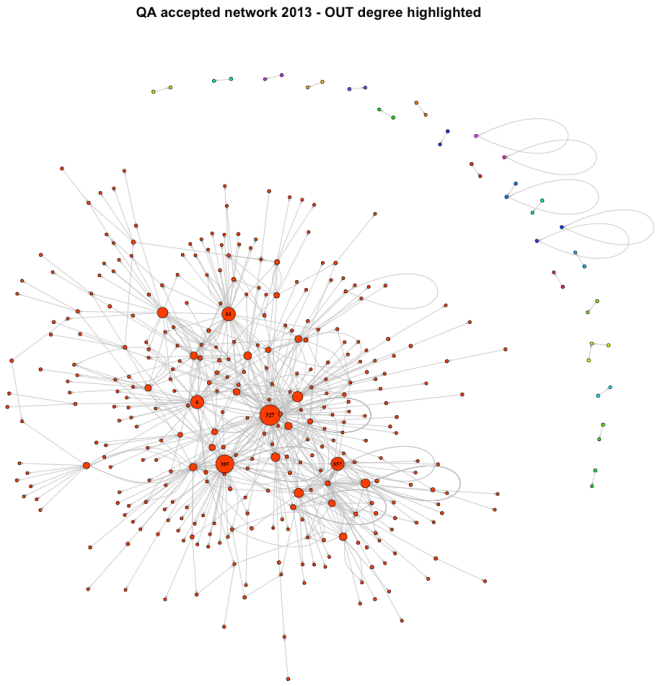

Figure 16 Network in 2013 (Out-Degree)

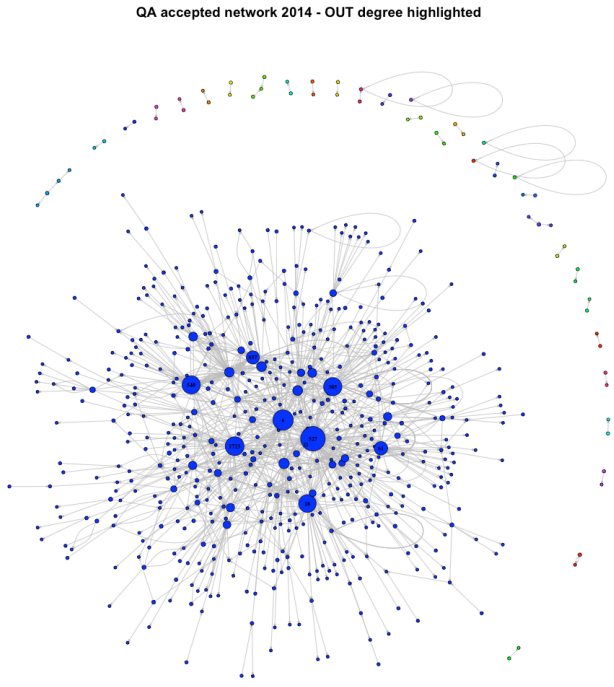

Figure 17 Network in 2014 (Out-Degree)

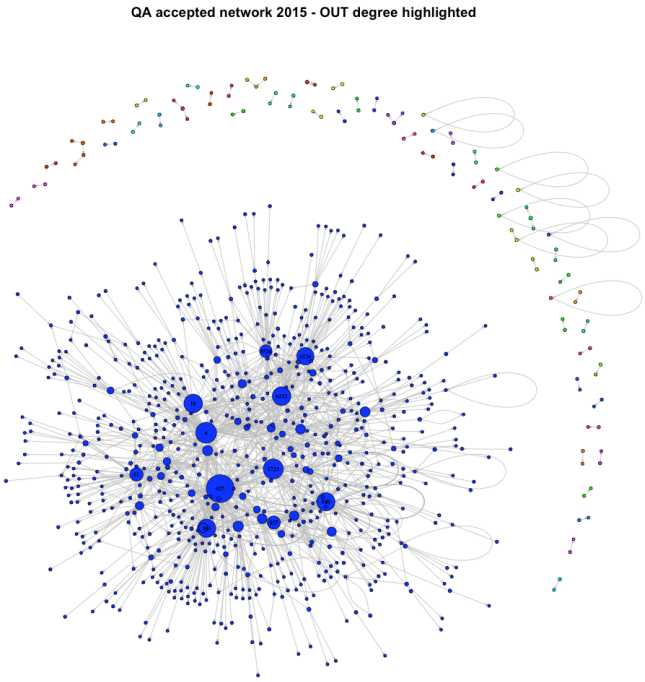

Figure 18 Network in 2015 (Out-Degree)

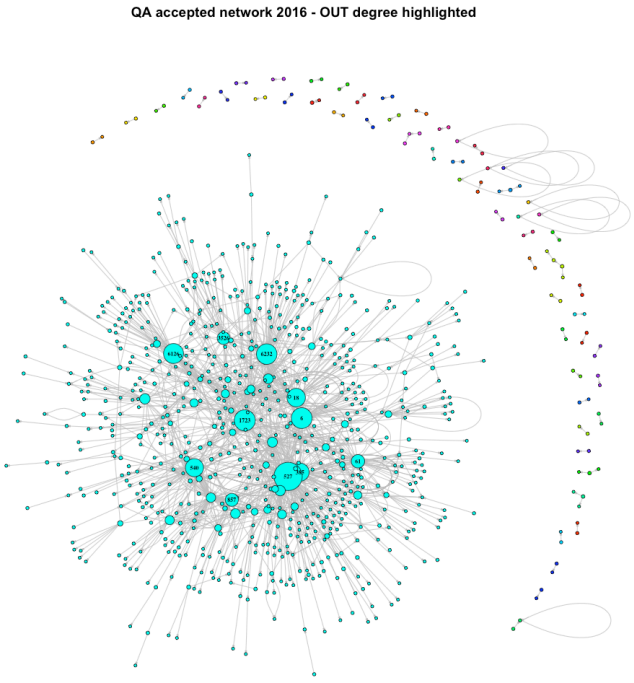

Figure 19 Network in 2016 (Out-Degree)

From all the visualizations above, we can obviously see that each year the network becomes more complex. Number of users (total circles) is increasing on each year. Number of questions with accepted answers (total ties) is also increasing on each year. Besides that, we can also see that there are more circles with larger size on each year (bot networks for in-degree and out-degree). This means that there are users with increasing number of accepted answers and answered questions on each year. In other words, the Q&A community is growing from the year 2012 until the year 2016. Therefore, we want to understand how do the network grows in details by looking at sample of users with 10 highest number of answered questions (in-degree) in 2016 and users with 10 highest number of accepted answers (out-degree) in 2016. Then, we look at years back at their degrees (both in and out) to understand how the numbers evolve. Bar charts below demonstrate the result of the analysis:

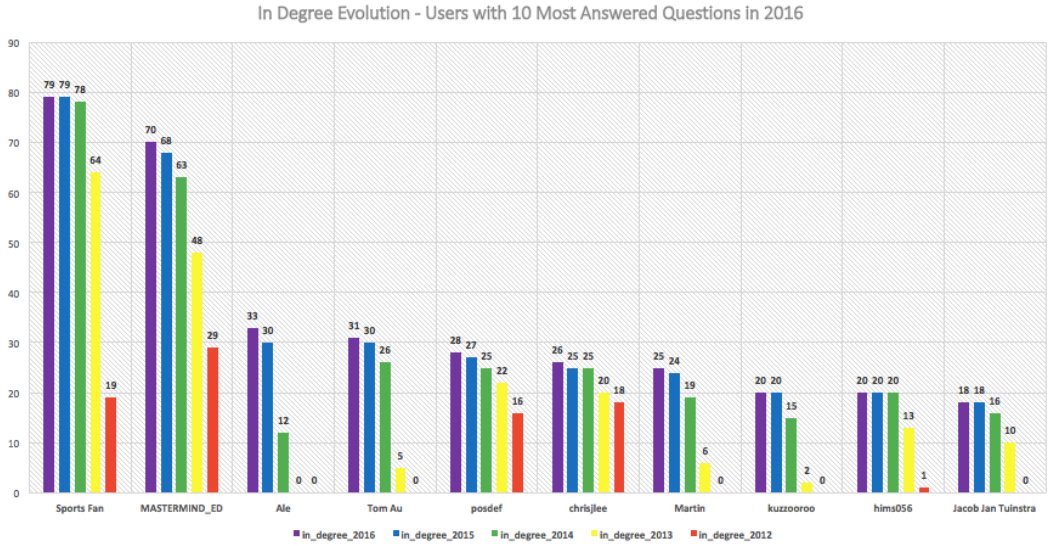

Figure 20 Top 10 Users for Out-Degree in 2016

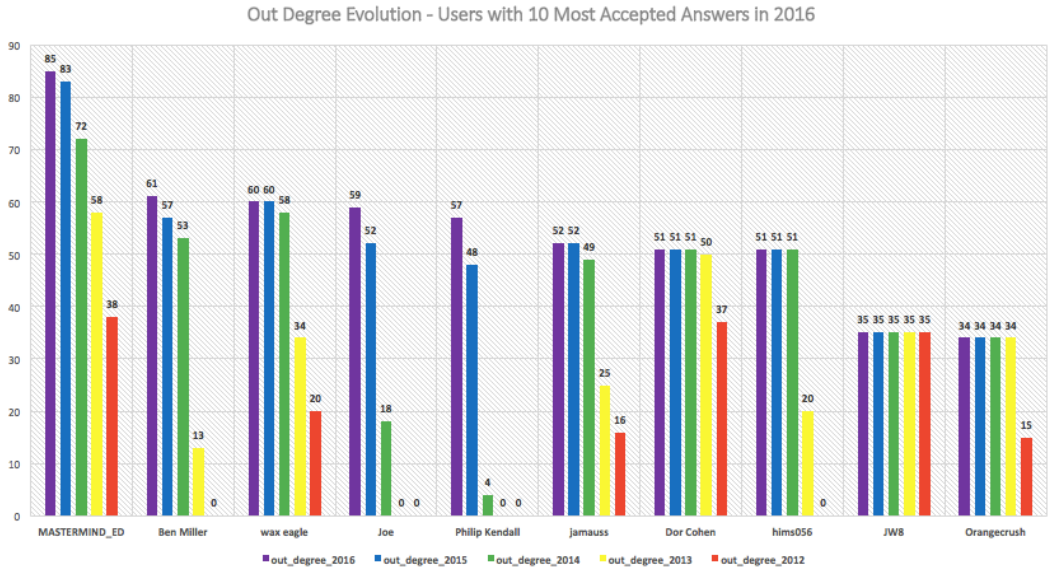

Figure 21 Top 10 Users for In-Degree in 2016

Based on the bar charts, we can see that some users that make it to the top 10 list in the year 2016, were not even existing within the network in the year 2012 (it is shown by zero value in the bar chart). Moreover, there are also users that were not always active (shown by increasing number of degree) who still make it to the top 10 list in 2016.

5 Conclusion

The interesting users are those who are always showing increasing out-degree or in-degree value on each year. We can infer knowledge that these users are keen to keep providing helpful answers and meaningful questions from year to year within Q&A online community. Thus, we can use these users as persona to understand what motivate them to do such activities. For instance, the insights can be gained by mining more detailed activities data from these users. Furthermore, the intervention can be in the form website's features customization based on the insights. Therefore, optimized growth of Q&A online community can be achieved. This same approach can be generalized to also optimize other topics in online Q&A community (not only Sports).

Appendix

Table **Users:**

Column Name	Description
_Id	Identification number of a row of this table.
_Reputation	Is a rough measurement of how much the community trusts you
_CreationDate	Date when the profile was created.
_DisplayName	User name.
_LastAccessDate	When user access the site last time.
_WebsiteUrl	Personal external website of user, can be null.
_Location	Where the user is located, can be null.
_AboutMe	Description of a user, can be null.
_Views	Number of times user's profile is visited by others.
_UpVotes	How many up votes has user's posts/comments received.
_DownVotes	How many down votes has user's posts/comments received.
_AccountId	Identification number of the account.

Table **Comments**:

Column Name	Description
_Id	Identification number of a row of this table.
_PostId	ID of post with which is comment associated.
_Score	Total score of the comment.
_Text	Actual text of the comment.
_CreationDate	When the comment was created.
_UserId	ID of user that wrote the comment.

Table **Votes**:

Column Name	Description

_Id	Identification number of a row of this table.
_PostId	ID of post with which is vote associated.
_VoteTypeId	Number that represents vote type, vote types are followings: 1: AcceptedByOriginator, 2: UpMod, 3: DownMod, 4: Offensive, 5: Favorite - if VoteTypeId = 5 UserId will be populated, 6: Close, 7: Reopen, 8: BountyStart, 9: BountyClose, 10: Deletion, 11: Undeletion, 12: Spam, 13: InformModerator
_CreationDate	When the vote was created.

Table **Tags**:

Column Name	Description
_Id	Identification number of a row of this table.
_TagName	Name of the tag.
_Count	Total score of the comment.
_ExcerptPostId	Post with tag excerpt.
_WikiPostId	Post with tag wiki.

Table **Posts**:

Column Name	Description
_Id	Identification number of a row of this table.
_PostTypeId	ID of post type - 1: Question, 2: Answer
_AcceptedAnswerId	Only if _PostTypeId is 1, what is the post id of the answer.
_ParentId	Only if _PostTypeId is 2, what is the post id of the question.
_CreationDate	When the post was created.
_Score	Total score of the post.
_ViewCount	How many users saw the post.
_Body	Text of the post.
_OwnerUserId	Id of user who created the post.
_LastEditorUserId	Id of user who edited the post last.

_LastEditDate	When the post was lastly edited.
_LastActivityDate	When there was last activity regards the post (comment/post/..)
_Title	Title of the post.
_Tags	Tags related to the post.
_AnswerCount	How many other posts react to the post.
_CommentCount	Number of comments related to the post.
_FavoriteCount	How many times was post marked as favorite.