

ASK MILLIONS OF PEOPLE EVERYDAY

Exploratory Analysis of asking and answering dynamics in Naver Jisik-In Question/Answer site

Kevin Kyung Nam
Cheng-Lun Li

*School of Information
University of Michigan*

INTRODUCTION

In recent years the online venues for creating and exchanging knowledge have greatly expanded. In the early days of the Internet, various topic oriented newsgroups supported their users by allowing them to freely ask and answer problems that users might face. Online forums that followed with the proliferation of the World Wide Web provided new opportunities for people to discuss many issues. The knowledge creating and sharing venues took yet another interesting turn with the newly created general-purpose and open-to-public question-answer sites. What is unique about these QA sites is that they do not limit themselves to one or few focused topics as typical newsgroups and online discussion forums do. Rather, it readily invites people with varying interests in different topics, creating a extremely large repository of “everything one wants to know”. These sites run by participating users’ asking and answering activities, with some moderation features and incentive structures. Some of the most well known sites include now defunct Google Answers and Yahoo! Answers. These sites have gained much research attention lately for their interesting roles as a knowledge creation and sharing facilitator. Not only that, the QA sites can be thought to support a new type of “search” method where a user let other users find an answer for him or her. And since many users get together and interact with each other through the question/answer activities, there are interesting possibilities for community forming and social networking within the QA site.

In this paper, we look at South Korea’s top QA site, Naver Jisik-In, and try to tease apart interesting aspects using a variety of methods, including social network analysis and quantitative and qualitative analyses. More specifically, we examine the following two areas:

- The network structure: we take a look at subsets of the question/answer activities from the site, and apply various social network analyses to study the structure.
- User participation behavior: we compare user participation behavior among different user types. We also study the incentive system and its effects.

The paper starts with the introduction of the Naver Jisik-In QA site, followed by a brief discussion on the data collection. Detailed findings in the network structure and user participation behavior analyses are discussed next, with a conclusion and future works.

NAVER JISIK-IN

Naver.com is South Korea's top search portal with rich features. According to a NY Times article¹ 77 percent of all searches in South Korea is performed through Naver.com, while mere 1.7 percent is through Google. However, web contents written in Korean were scarce in the early years of the Internet, and the QA site Jisik-In was created, in part, to create searchable contents to attract traffic to the portal site. "Jisik" in Korean means "knowledge". "In" has the same pronunciation as the Korean word "person" or it can also be interpreted as the English word "in". And thus the name of the site can mean both a "knowledgeable person" and "knowledge-in". Either case, the name is clearly indicative of the site's goal to create, garner, and share knowledge among people.

The QA site has grown explosively since its first service in 2002. According to Naver.com's claim, its database contains over 72 million questions, with 16 million visitors and 44,000 new questions per day¹. To put this into perspective, the entire South Korean population is around 51 million, of which about 34 million are the Internet users.² It is estimated that there are 6 million to 12 million registered Naver users, so the stats clearly indicate that the site is very active. The QA site has 13 top categories, each further branching out two or more levels, totaling up to more than 3,500 categories. Some categories are cross-listed. For example, the "Cyber culture" category is listed both in the *Computer/communication* top category and *Social/Culture* top category due to its relatedness.

A number of incentive features are implemented in Knowledge-In to encourage user participation. The most obvious and visible one is the point system that creates a type of closed economy within the Naver portal sites. There are many ways a user can earn points; initially by signing up for an account, logging into the system each day, performing various moderation work within the QA site such as rating (recommending) others' answers, and of course by answering a question and being selected as the best answer. Points are of no use if one cannot spend it. A user can spend the earned points to ask questions and post reward points, purchase items the user can use in the Naver site, download documents such as publications and articles. Points not only decide the purchasing power of a user, but they also act as a "score" for the user's overall activities in the system. There are fourteen elaborate ranks a user moves up as he or she accumulates points through activities. Each rank is associated with an aptly named flamboyant avatar, clearly indicating the user's "social status" within the site. Visibility of a user's id is also determined by the points he or she has. The top thousand users are listed on the

¹ Choe, Sang-Hun. "South Korean Connect Through Search Engine" *The New York Times* (2007): 1 page. 05 Jul.2007 <

²KIUSPSE. *Korea Internet Word Stats and Marketing Report*. 2007.

<<http://www.internetworldstats.com/asia/kr.htm>>

main site where everyone can see, and this may give some people a reason to be active in the site. One flaw of the way the point system is implemented is that you can fall to the lower rank if you spend enough points. Thus, it hampers some users' willingness to spend points by asking questions or purchasing services, and makes them "point hogs", which may have a negative impact on the overall site activities.

A user can post a question in the appropriate category and other users who are willing to answer would take up on the task. There is a type of backchannel where you can directly ask a certain user, but we did not consider this data in this study because the data was considered a private communication between users and was unavailable for collection.

DATA COLLECTION

Since the site contains a huge dataset and its public API only allows access to a limited amount of data, we have created a couple dozen parsers in Perl to crawl and collect different parts of the QA site. (Most of the parsers had to be rewritten toward the end of the project after Naver changed the format of its site.) There is a limitation to how far back you can see the questions that are listed in the site; each category lists only one hundred pages with thirty questions per page. (This has changed in the later new format.) For each bottom category (the lowest level category), we collected one hundred pages of questions. Since each top category has multiple bottom categories, we could acquire relatively big data set for our analysis. The user information page for each user lists the entire history of his or her question and answer activities, so in theory one can gather most of the site's data by finding all the users and harvest data from each user's information page, albeit this will require more computing resources and time. Data collected using the parsers are then stored in appropriate tables in the MySQL database. Among the collected data, the following list shows what is actually used in our analyses:

- Question/Answer entries for two top categories
 - o Computer/Communications: 85,327 questions and 139,434 answers
 - o Entertainment: 67,603 questions and 142,250 answers
 - o User information from the above two categories: 182,729 users
- Question/Answer entries for top 872 users
 - o 2,694,397 entries
- Rated Question/Answer entries for top five categories

	Num question entries	Num answer entries	Num user information
Business	5,492	13,384	3,924
Computer	5,124	8,768	3,753
Entertainment	5,494	26,119	16,961
Health	5,465	15,224	7,425
Shopping	5,485	12,816	6,916

Overall, we have collected over 1 GB of textual data.

NETWORK STRUCTURE ANALYSIS

General structure

Comparison to Java forum

Bow tie structure

Given the nature of the QA sites, we expect the network structure of the question/answer activities to be very sparse. As Zhang et al claim, the Java Forum in their study presented a very different bow tie structure than that of the World Wide Web. The largest strongly connected component (LSCC) of the Java Forum was much smaller than that of the Web. However, we found that in Naver QA site, the LSCC was extremely small (32 nodes). As we can see in the following table, the IN and OUT of the two categories in Naver are also very small. We can say that there is not much discussion among users within these two categories, rather there is directionality in who asks whom. Moreover, there are many one time askers and answerers in both categories. For instance, in Computer category, 55% of the users only ask and 44% of the users only answer questions. Similar situation is observed in the Entertainment category (only ask: 42%, only answer: 56%), too. Another interesting result is the large percentage of the tendrils. In both categories, tendrils consist of over 75% of the entire structure. It seems a relatively small number of answerers in In and Out parts of the bowtie answer a huge number of questions asked by people who do not ask among themselves.

	Core	In	Out	Tendril	Tube	Disconnected
Web	27.700%	21.200%	21.200%	21.500%	0.400%	8.000%
Java	12.300%	54.900%	13.000%	17.500%	0.400%	1.900%
Computer	0.032%	0.049%	0.100%	76.119%	0.049%	23.650%
Entertainment	0.159%	2.338%	1.866%	69.609%	0.252%	25.775%

Table 1: Bow tie structure of the web, Java forum, and two categories in Naver

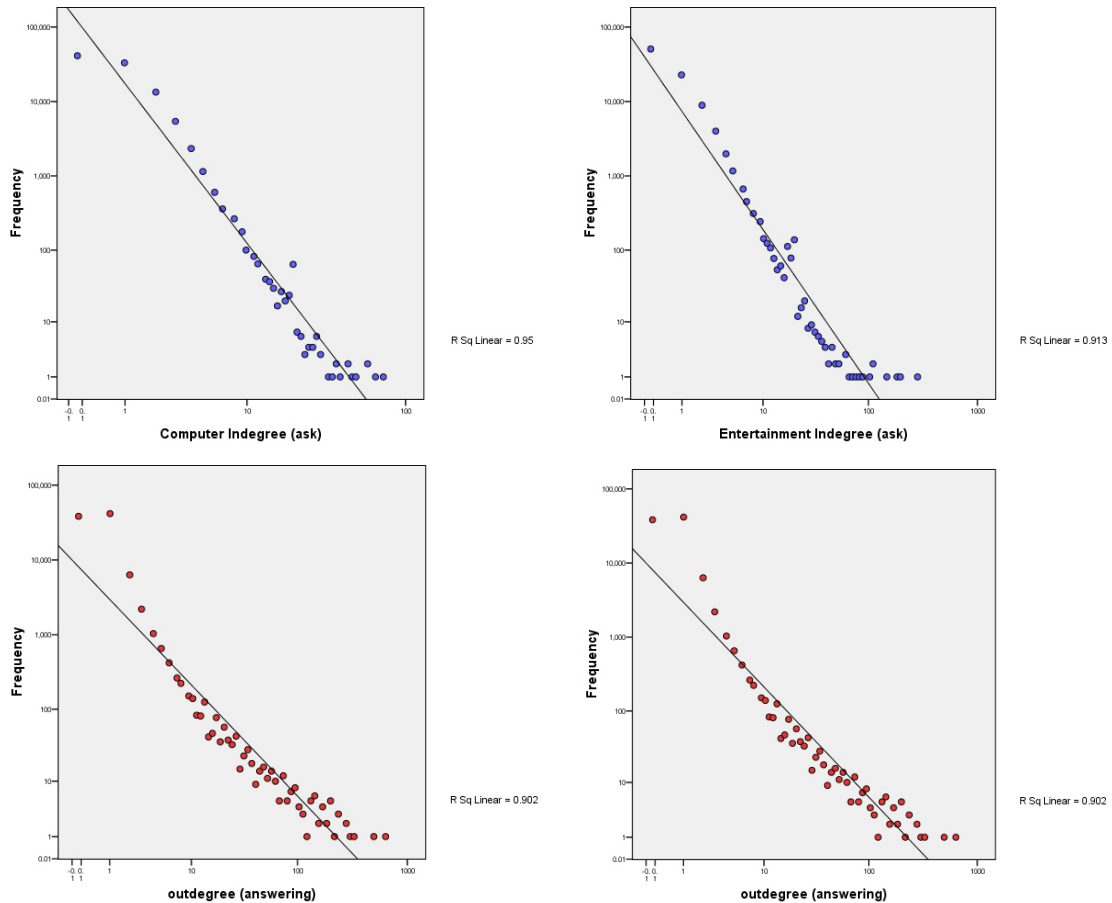
Comparison between the categories

It is clear that the structure of Naver differs much from the Java forum and the web. We next compare the categories within Naver. Do they have different network characteristics? Do the users in different categories behave the same way? We picked two categories, Computer and Entertainment related categories, and compared them against each other. These two categories were picked partly due to our conjecture that the required expertise levels in these categories would be different and thus exhibit distinguishable user behaviors. We hypothesize that the computer related questions would require more expertise level than the entertainment related questions. To start, we first look at some of the centrality measures, such as density, and degree distribution.

	Computer	Entertain	Both categories
Average Degree	2.36	2.29	2.34
Density	0.0000120	0.0000124	0.0000068

Table 2: Density of Naver categories

It seems that the average degree and density of both categories have similar values. The first finding we have from the plot is that both categories demonstrated a power law distribution. This is common in QA sites. In QA sites, most of the users ask/answer very few questions, while a smaller number of users ask/ answer a lot.



Lo-log plot of degree distribution of both categories

In order to get a better sense of how skew the distribution is, we picked top 100 users from both categories and checked how many answers among all they answered. We found that even if the 100 users consist only **0.2%** of the users participated in computer category, they contributed **16%** of all the activities (mostly answering). Similar situation is observed in the entertainment category.

If we look at the sub-categories of Naver, we can see the activity level and the density of the answerers varies a lot in different sub-categories (**Figure 1**). “Total messages” shows the popularity of the subcategories. “Answers per questions” show how many answers each questions has. “Answers per answerers” shows how active answerers of these subcategories are.

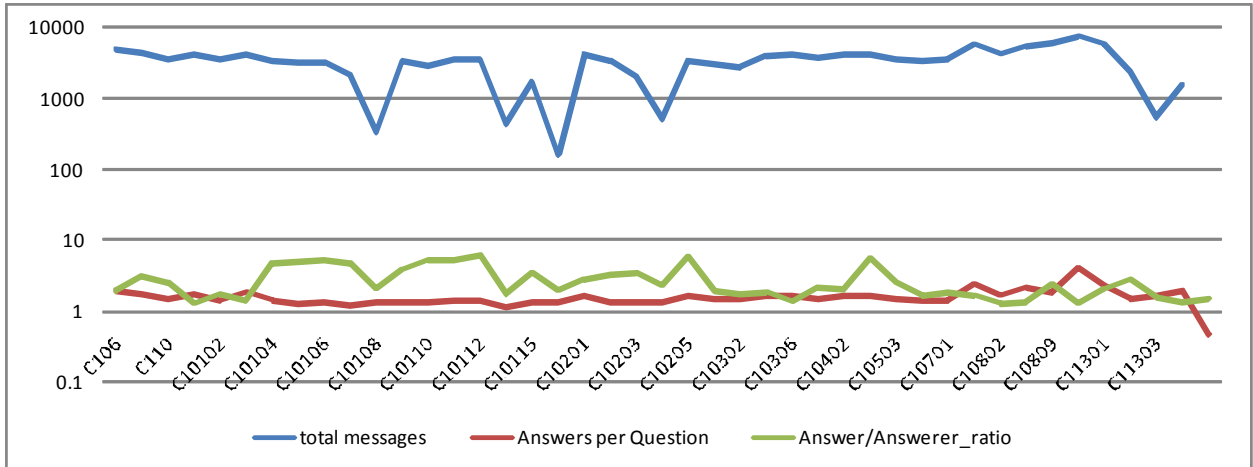


Figure 1: Answers per question (red), total answers (blue) and Answers per Answerer (green) of computer category

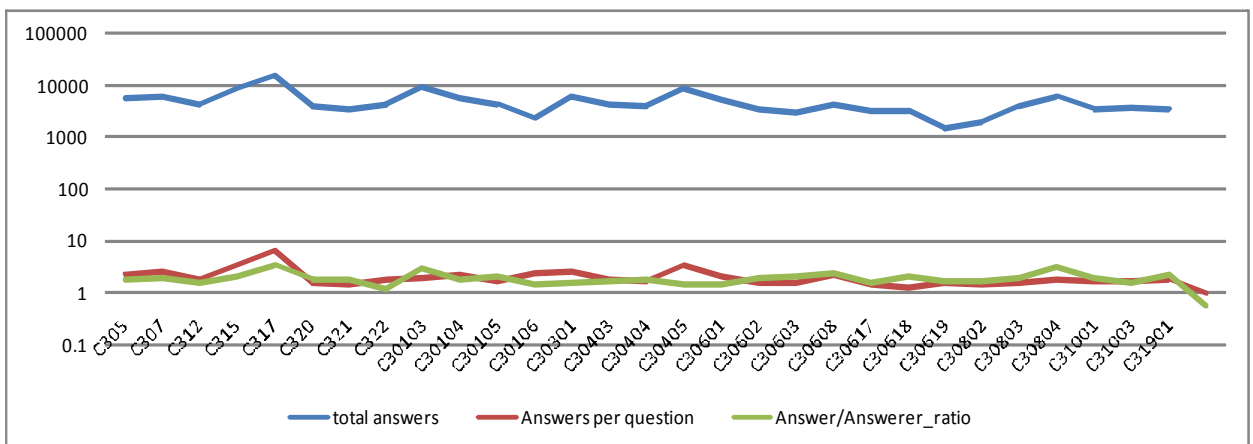


Figure 2: Answers per question (red), total answers (blue) and Answers per Answerer (green) of Entertainment category

There were some interesting phenomena we noticed in the plots:

- 1) Popularity of categories: Generally speaking, subcategories of Entertainment are more popular and had more questions (mean: 4907) than Computer (mean: 3322). If we take a closer look of the total number of answers, we can tell that there is more variance in computer categories than entertainment.
- 2) Answerer density: Ideally, it would be better for each question to be answered by more users. Looking at the number of answers per answerer, we found the variance in computer in much higher than entertainment. This means in some categories, questions are expected to be answered by more users while others are not.

Motif structures

Since the asker-answer networks in both categories contain many nodes and edges, it is hard to make sense of the local structure in them. In order to examine the local structures, we conducted a motif analysis. Motif analysis helped us understand the basic components of the whole network.







						
Computer	89.1%	10.67%	0.21%	0.02%		
Entertainment	96.67%	2.98%	0.28%	0.02%	0.03%	0.02%

Table 3: Motif structures for both categories

Since the out-degree means the nodes (users) answer others' questions, we can see that in entertainment related questions, questions are answered by more answerers than computer related questions. The result is reasonable because there are more questions in the entertainment category that generate a "discussion", rather than a straight-forward answer. (For example, "what is the best show in the 90s?")

Role of top users

As mentioned above, a small number of top users generate a huge portion of the answers in the Naver QA site. Does that mean the site depends on these few users? What if we take these users away? How much impact will that bring to the network? Will the site still be functional? In this section, we will focus on the difference in structure and the efficiency (the response time and number of answers per question), with and without the top users.

We picked top 100 users in contribution and remove them from the asker-answer network and analyze the difference in three major topics of interests: 1) response time and 2) answers per question.

Impact to response time

We compare the difference in response time in both categories before and after the top users are removed. The result is:

	Computer	Entertainment
Original network	25.13 hours	29.24 hours
Top 100 users removed	28.32 hours	32.42 hours

Table 4: Response time before and after top 100 users removed

We can observe that about 3 hours are added to the overall response time after removing the top 100 users.

Impact to answered questions

And we try to find out how many questions remain unanswered after removing the top 100 users:

	Computer	Entertainment
Number of questions become unanswered	8649	21549
Percentage	12.8%	25.3%

Even if we remove only the top 100 users, they have already caused significant effects to both categories. However, we do not claim this result would be reproducible in the original network

in the absence of the top users, since already answered questions would attract fewer answerers and unanswered questions would attract more users' attention. The impact of the top users is still significant, and they are nonetheless important members of the site.

We use an alternative way to examine the effect of the top users by removing them from the network until the average number answers per question drops below 1. We find that for the computer category, the answers per question ratio drops below 1 after about 1,050 (2%) top users are removed. For entertainment category, the number is about 9,000 (14%).

PARTICIPATION BEHAVIOR

Incentive structure – point system

In this section, we examine what effects the point system has on users' question/answer behavior. When a user posts a question, he or she may post extra points up to 100 which will be given to the selected answerer. An answer provider receives two points for leaving an answer of any quality, and 10 more points if selected by the asker as the best answer plus whatever points the asker posted extra. Intuitively, a question with high points would attract more viewers, perhaps a quicker response and more competition among the answerers. A number of previous studies found that there is little correlation between the reward and result. Our data shows the response time (the time it takes for the first answer to appear) is a little bit shorter and number of answers higher for questions with higher points, but the effect is not consistent among the categories. **(Figure 3)** We have used the "recommended questions" section of the QA site to collect around 5,000 questions in each of the *Business*, *Computer*, *Entertainment*, *Health*, and *Shopping* category. The recommended questions are ones that have at least one user rating (a thumbs-up). Before Naver changed the format of the QA site, there was a section named "Completed questions", where all the questions that were deemed completed either after the asker chose the best answer or a certain period time had passed for were listed. After the format change, however, Naver changed their policy on how to assign a status on a question and completely removed the section. The recommended questions section was the closest to the former completed questions section, with at least one answer provided to each question. After data collection, we calculated and plotted the response times and average number of answers. **(Figure 3)**

A few observations have led us to conjecture that users are driven more by the efforts one has to exert to provide an answer, popularity of the topic, and points if everything else is about equal. This is subject to a further study and we do not provide any quantitative evidence in this study.

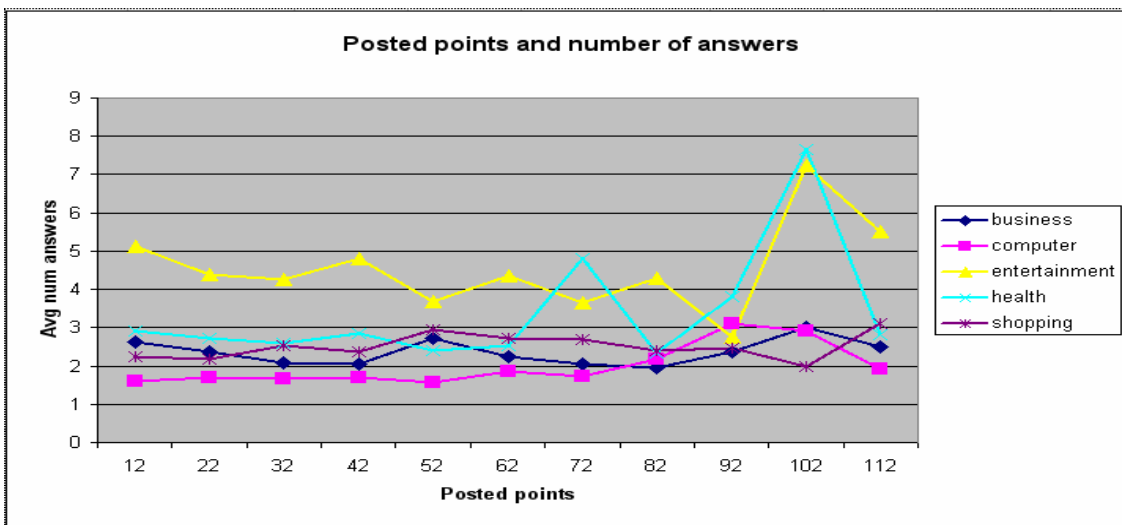
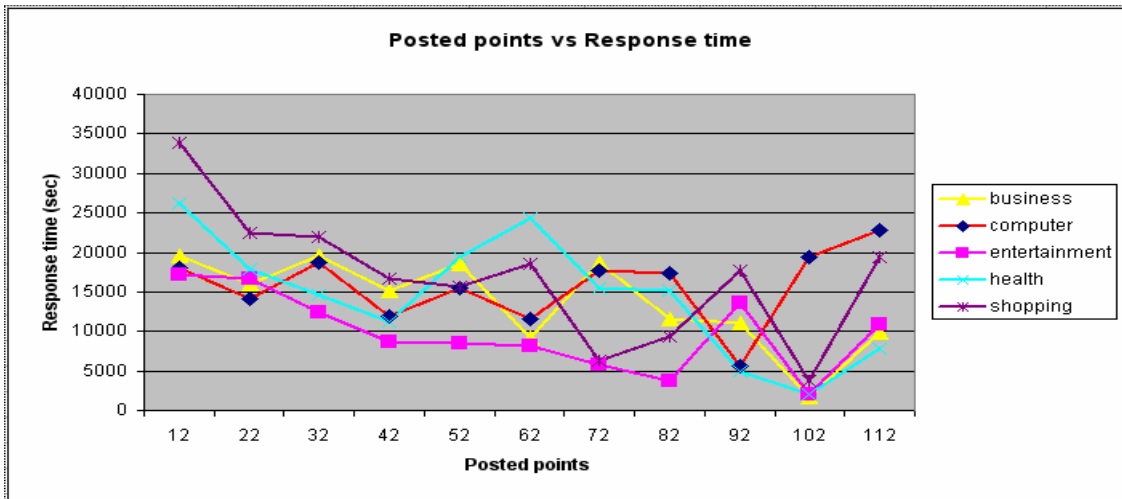


Figure 3

Comparison between top users and other users

As shown in the network structure analysis section of the paper, the QA activities follow a scale free network where there are few users who answer many questions while most of the users answer one or a few. It is intuitive to think that the most active users would behave differently than the average users. In this section, we examine what differences the top users and other users exhibit in our data.

Data collection:

We collected participation data of the top 1,000 users that are listed on the site. The rankings are decided by the amount of points users have accumulated, and since the points can be earned in many different activities including answering users' questions, the top 1,000 users are not necessarily the most frequent answerers. In fact, a couple users maintain a strange status of being within the top 1,000 with answering no question. There must be users with more answers

below the 1,000 people (some observed in our randomly selected user data), but there was not an easy way to collect this data. After only selecting users with sufficiently many answers, we examined 872 top users, whose answers totaling up to 2,694,397. We also used users from the two categories (Computer/communication and Entertainment) from the network analysis section as our average users. (Excluding the top users from the categories)

	Avg answers	Avg questions	Best answer ratio	Avg recommendations	Avg recommendation per answer
Top users	3090.0	29.3	73.7 %	3494.5	1.13
Other computer users	298.2	21.1	59.8 %	119.0	0.40
Other entertainment users	292.2	25.9	56.5%	131.0	0.45

Table 6

Discussion:

It is clear from the data that top users consistently leave better answers on average, and rated better by other users. The ratio between the number of answers and questions differ much between the top users and others. The top users ask only once in every 100 answers they provide, while others are about once in every 10. These users are indeed a valuable asset to the QA site. They not only answer a lot of questions, but other users view their answers as of high quality. However, our data does not reflect whether the top users are necessarily “experts” in the domain. They may have answered easy questions with good explanations, (thus leading to many “thumbs-up”s), but nothing reveals about the difficulty of the questions. This is subject to a further study in which a content analysis would reveal the answering behavior better.

User motivations according to level

What motivates people to participate in the QA site? For the askers, it is rather obvious. They have some information needs, and the QA site can be one way to get the solutions. The motivations for the answerers who put efforts into finding questions they can answer and writing up solutions are of great interest. In this section of the study, we try to compare possible motivations for the users expressed in their comments left at each level-up. As a user accumulates points through various activities, he or she moves up to the next level. From level 6, the user is given a chance to leave a comment often consisting of thanking everyone, expression of his motivations, and future plans. What is interesting is that the motivations for participation expressed in their comments differ between some of the higher level users and lower users. Here we discuss our findings from their comments.

Data collection:

Not all comment reflects the user’s motivations. Of those that contain an explicit statement of motivation and implicit but strong evidence for a motivation, we have collected 184 comments. Since there are fewer users at the highest levels, we have examined every comment from level 14 to 9. There are a few thousand comments at the lower levels, so we randomly picked comments for those levels. We used the following guideline to categorize each comment. A comment may have more than one category.

- **Altruism:** display of user's purpose as helping others, sharing knowledge
- **Level-up:** display of specific goal oriented reasons, including leveling up, change of avatar status, stat improvement
- **Self-growth:** display of self-growth or learning as primary reason for participation

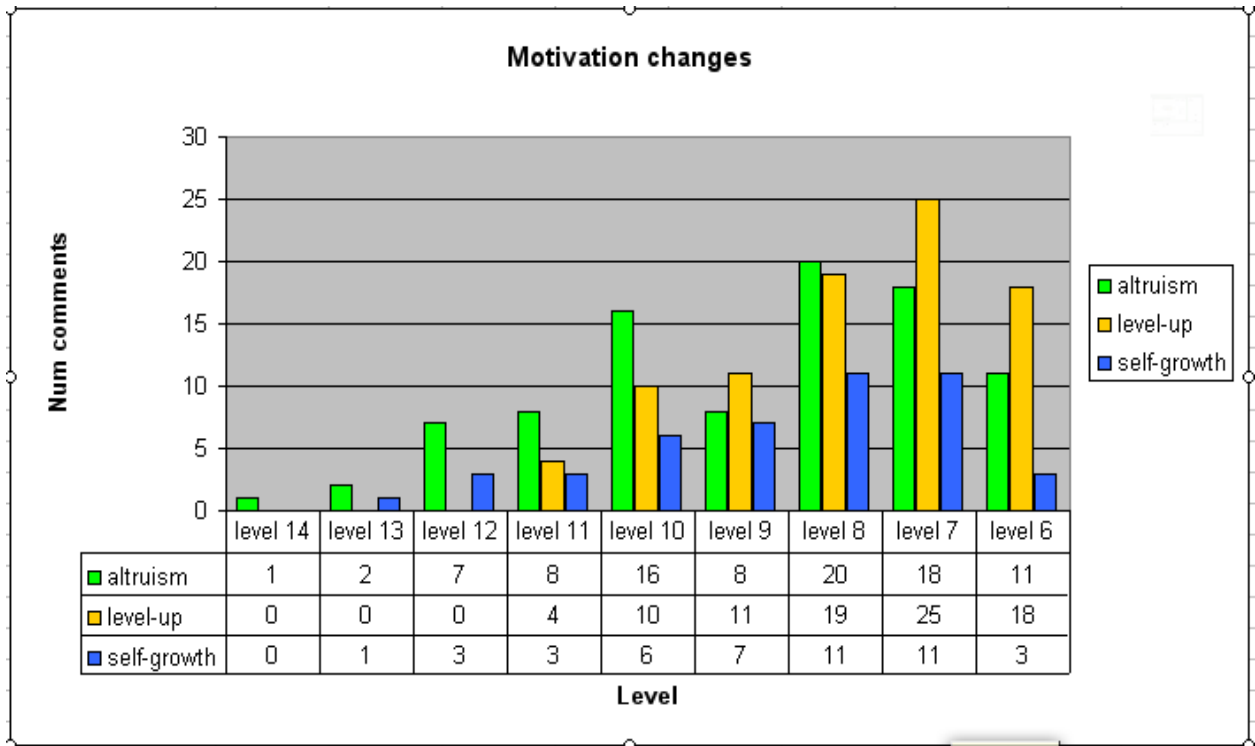


Figure 4

Discussion

What we have found clearly shows that motivations for users are different depending on their levels. Users at level 10 and above state altruism for their primary reason for participation. There are very few comments of level-up as their goals. More users at the level 9 and below state level-up as their primary goals as well as altruism. The implication may be that the site needs to implement different incentive mechanisms for people according to their experience level.

Are the top level users fundamentally different people from those at lower levels? Or people's motivations change as they stay longer in the system? There are only few top users who have consistently left comments along their paths to the top, and thus we do not have enough evidence to generalize it in one way or another. Four top users have left about 5 to 6 comments along their paths, and it seems three of the four consistently have stated their motivations as helping others, while one user has stated that his initial motivation was collecting points but later changed to helping others.

Do users really mean what they say in the comments? Even though one may state and really feel that helping others is the main motivation, the user may still be driven by other reasons and his or her answering behavior may reflect this. To understand whether their stated motivations are

reflected in their question/answer behaviors, we have collected the entire history of a mix of users that expressed their motivations strongly one way or another.

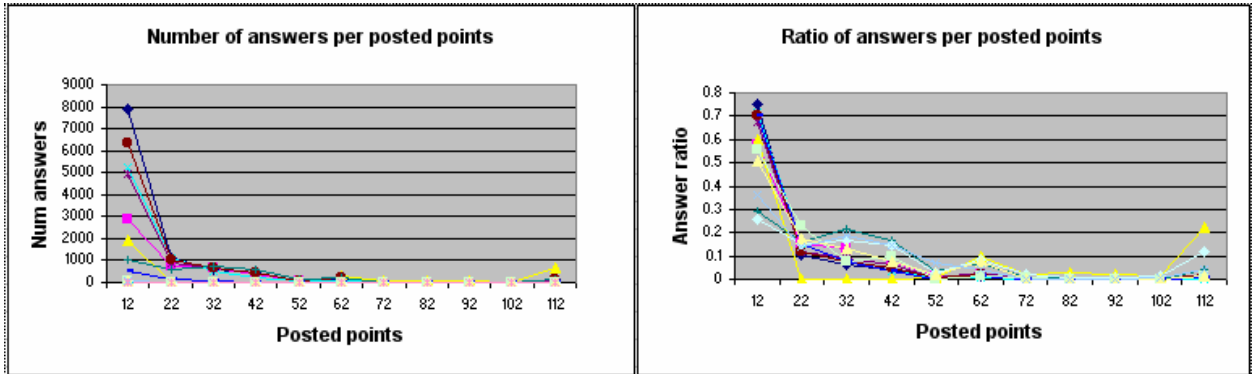


Figure 5

The number of questions per posted points that users answered were not very different between the motivations. The ratio of the questions per posted points within each user's answer pool was more distinguishable, but there was no consistent pattern in the same motivation that can be generalized. We then looked at the answering history of each user. (with the longest participation lasting about 5 years) For each month, we plotted the number of questions each user answered and the average posted points of the answers.

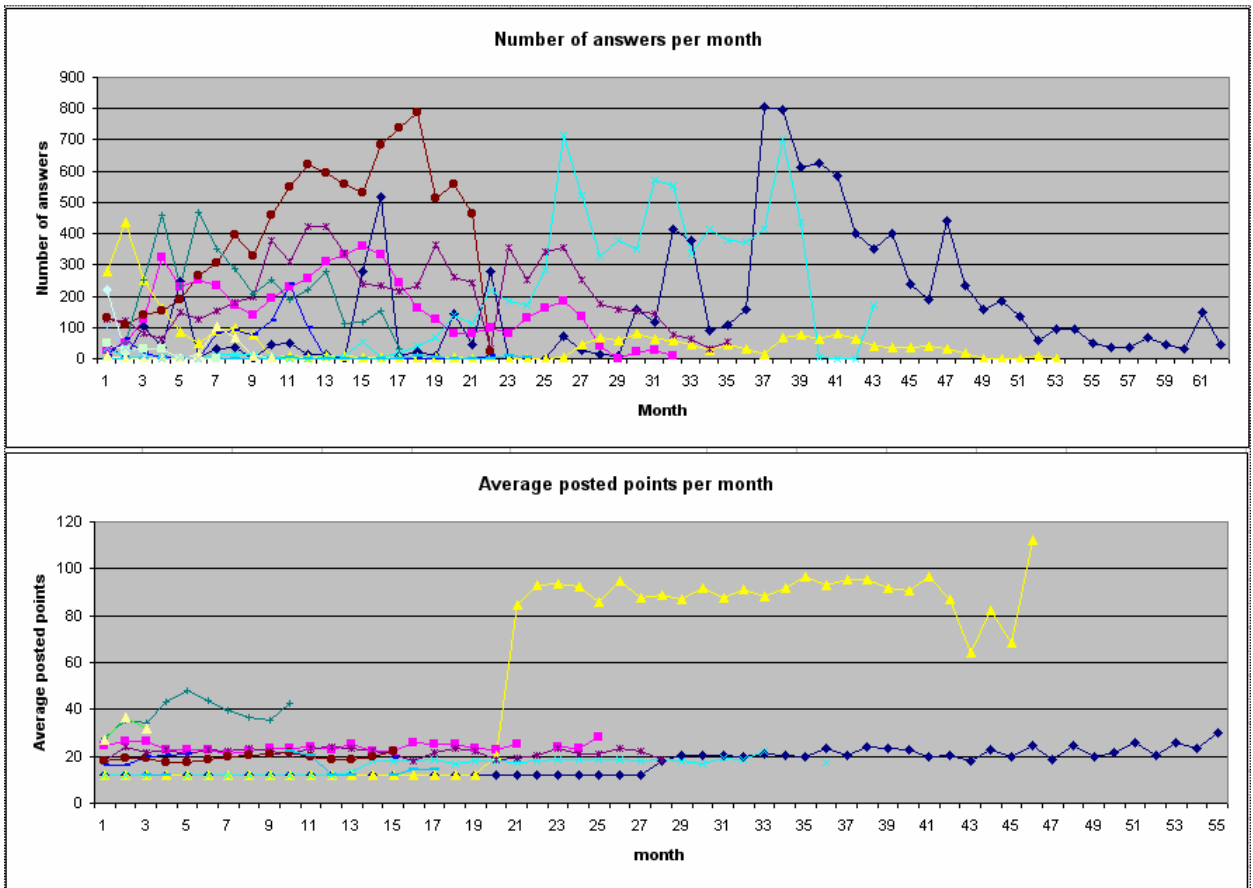


Figure 6

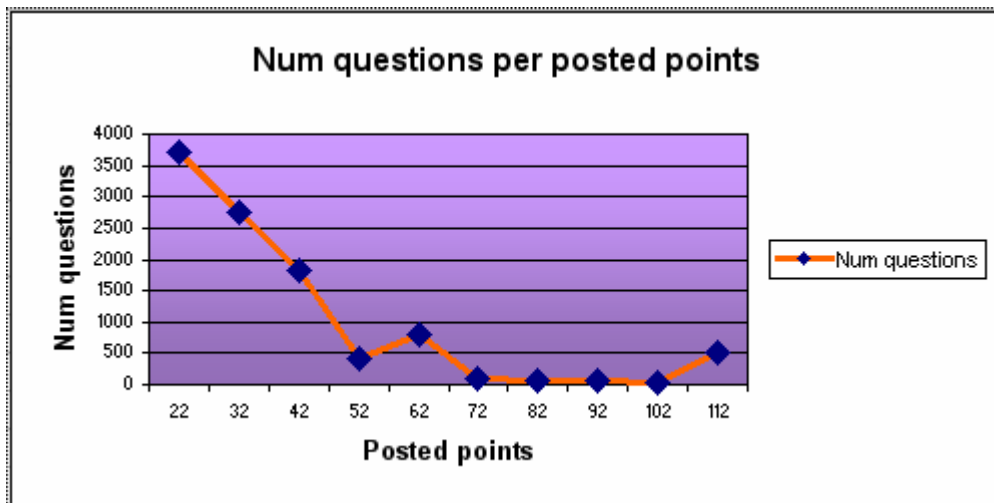
Again, the number of answers or average posted points that each user left in a month did not show any recognizable pattern within the same motivations. However, there was one noticeable user (represented as a yellow line with a triangle indicator) whose behavior changed much after a year and half. The user mostly answered with very low point questions (12 points) during the earlier year, but suddenly took only higher point questions afterwards. We have looked at the user's comments, but unfortunately he only left two comments which did not provide any clue to what had contributed to his answering behavior changes. A further study may be needed to look at other places that may reveal user's motivations better.

Askers' valuation of their questions

How do askers decide the amount of extra reward to post when they ask a question? From reading the actual questions, it seems users post high points when they feel 1) the question is difficult, 2) the question may take long to answer, and/or 3) a quick response is desired. But do the askers judge correctly the cost of answering their questions? In this section, we present initial evidence that they do not. While this evidence is limited and it requires further scrutiny, we feel it provides at least an insight to how well users are aware of their questions' costs.

Data collection:

We have randomly collected 27,716 questions in 612 categories with 64,438 answers from 28,325 answerers. We then calculated how many questions exist for each reward point interval. (We exclude the default 12 points, i.e., no extra point posted.)



Discussion:

It makes sense economically that there would be many questions with smaller extra points, while few with the highest points. Thus, intervals between 22 and 52 more or less follow what one can expect to happen. An interesting part is between 62 and 112. It clearly shows that users do not distinguish much between 72, 82, and 92. Their questions are only a few dozens compared to the 28,000 other questions. However, there is a sudden increase at 112, the highest points one can post, and this may be indicative of the fact that users may think in terms of small points and large points, rather than in a finer granularity of the points.

CONCLUSION

In this study, we have shown our exploratory analysis of the Naver Jisik-In QA site. The network structure clearly shows that the site follows a scale-free network, where a handful of users answer many questions while most other users answer few. Users' question/answer behavior differs among the topic categories, and top users are distinguishable in their activities in the site. However, further study is needed to truly reveal the question/answer behavior of the users. The current implementation of the Naver QA site does not allow much social networking among the users. There is little direct interaction or discussion among the users, and community formation barely exists. It would be interesting to examine what features may improve the quality and the usefulness of the site.

REFERENCES

Zhang, J., Ackerman, M. S., and Adamic, L. Expertise Networks in Online Communities: Structure and Algorithms, *International World Wide Web Conference (WWW2007)*, Banff, Canada, 2007.

Welser, H. T., E. Gleave, D. Fisher, and M. Smith (2007). Visualizing the signatures of social roles in online discussion groups. *The Journal of Social Structure* 8 (2).

Zhang, J., Adamic, L. Bakshy, E. and Ackerman, M. Everyone Knows Something: Exploring Knowledge Sharing in Yahoo Answers, working paper (submitted to www'08).

Choe, Sang-Hun. "South Korean Connect Through Search Engine" *The New York Times* (2007): 1 page. 05 Jul.2227

<http://www.nytimes.com/2007/07/05/technology/05online.html?ex=1341288000&en=9610e1983f36f1e6&ei=5090&partner=rssuserland&emc=rss>

KIUSPSE. *Korea Internet Word Stats and Marketing Report*. 2007. <http://www.internetworldstats.com/asia/kr.htm>