Modeling huge social networks changes using the topological change of the most active players’ networks

Amit Rechavi
The Center for Internet Research
Graduate School of Management, Univ. of Haifa
Haifa, Israel
Amit.rechavi@gmail.com

Sheizaf Rafaeli
The Center for Internet Research
Graduate School of Management, Univ. of Haifa
Haifa, Israel
Sheizaf@rafaeli.net

Abstract

This study investigates the active users in social networks. First we investigate the importance of active users (as we defined them) in their community – a content category. Secondly we used the activity log of the most active users to build an artificial network. Lastly we explore the possibility of using this small data-set of intensive users to depict and explore the dynamics of a huge social network. We study this network and its basic topological parameters and find them to correlate with the overall activity volume of the entire social network.

Ten years after Barabási et al.’s (2002) study on real-life networks, our on-line Q&A social network study, reaches the same finding and conclusions regarding the connection between topology and network’s size. Our finding has positive implications for the investigation of huge social network.

1. Introduction

Online Social Networks (OSN) share, organize and search relevant content. Social Question and Answers sites (SQA) are Web-based information-seeking services in which questions are asked and answers provided by the users. These sites attract active and consistent users (Shah et al., 2008) and the answers from these sites have a higher quality than those of specialists (Harper, et al, 2008).

Yahoo! Answers is the world’s largest question-answer system and it acts as a community site (Harper et al., 2008) and as an “online social network” (Agichtein et al., 2007). Online groups are most successful when a leader sets the agenda and the quality of the answer (Kerr, 1986). In Yahoo! Answers there is no official leader or coordinator and the continuous and consistent users are those who contribute to Yahoo! Answers’ success (Shah, Oh, and Oh, 2008).

In this study we explore these continuous and consistent users in Yahoo! Answers and w call them: “active users”. These users respond to questions, create value, personify site norms, earn social capital, and have a strong influence on answer quality assessment (Gazan, 2011). These “user’s activities” (their precise definition is given in section 5) are of a special interest to the authors and in this paper we are interested in studying the following empirical questions:

(1) Existence and Importance. Do “active users” exist in all online social networks and what is the relationship between their activity and the network’s overall activity?
(2) Modeling. Is it possible to use these users and their activity only, to model the activities of a huge social network?

The paper has eight sections. We first review related work in Section 2. Then, in Section 3, we present the active users’ network. Section 4 includes our two hypotheses. Research methodology and the research results are reported in Sections 5 and 6 respectively. Next we discuss the results in Section 7. We conclude the paper with study limitation and future work in Section 8.

2. Related Work

2.1. *Yahoo! Answers*

Launched on July 5, 2005, Yahoo! Answers enables participants to ask and answer questions on any topic. It provides more than 20 million answers per month and it serves many needs - answering questions, receiving support, and requesting everyday advice (Adamic et al., 2008). Yahoo! Answers has more than twenty top categories and more than 1,600 sub-categories. Some categories are huge with more than 100,000 users per month and some have only a few users per month. Its content includes informational questions and conversational discussions (Harper et al., 2009) which are composed of opinion type questions, evaluation, and points of view (Kim et al., 2008). It provides good answers for conversational questions (Liu and Agichtein, 2008) and for categories where the answerers were active on specific topics only (Adamic et al., 2008).

Yahoo! Answers uses the crowds to get data and information. Unlike other social platforms such as Twitter, where the users must direct their question to a certain user to get results (Nichols and Kang, 2012, Paul et al., 2011), in Yahoo! Answers, posts are archived and can accumulate comments and views over a long time (much longer than a twitter transaction).

The *Yahoo! Answers* process is quite straightforward. An asker places a question on *Yahoo! Answers* by selecting a category and entering the question subject (title) and, optionally, giving details (description). The question is in an "open" state in which answers are received from the all users. Once the asker is satisfied with any of the answers, he or she can choose it as a Best Answer (BA) and provide feedback (e.g., stars or textual feedback). If the asker doesn't indicate a BA within four to eight days, the BA is chosen by the community in a voting process. Once a BA is chosen, the question is "resolved" and comments can be added to both closed and resolved questions. During the process users can tag other users' questions (e.g., award stars for quality) and answers (thumbs up or thumbs down).

2.2. Active Users in Q&A Sites

"*At the heart of small cliques are a few strong relationships, and as long as these persist, the community around them is stable*" (Palle et al., 2007).

The importance of the loyal consistence users of a community is a long studied issue. Morgan et al. (1996) suggested that at the center of any network are its stable "core" members, whereas "latent" vertices are peripheral and less stable. The core might comprise as little as 10% of the largest number of connected high-density vertices (Zinoviev, 2008) but once the core is removed, the social network disintegrates and loses its ability to function as a whole (Mislove, 2007). The core of the network can determine the group's nature (Backstorm, 2008) and the intensity with which team members participated in the team (Koh et al., 2007).

In this study, the active users are those who ask and answer consistently during a long-enough time period. The definition is a mixture of both the number of contributions, their quality (tagged as "Best Answers") and persistence (Lapas and Terzi, 2008) is critical too.

Our definition resembles the "Answer People" (Turner et al., 2005; Welser et al., 2007; Wellman, 2009) and "Most Active Users" (Yang et al., 2010). The active users are the core of the online social network. Their participation and stability are critical since participation and activity in social networks are rarely stable and 80% of the nodes appear in fewer than two snapshots of a social network (Bouguessa et al., 2008). The active users are indeed numbered, but they reach a high level of participation and account for the majority of the action (Lakhani and Von Hippel, 2003; Soroka and Rafaeli, 2006; Murata and Moriyasu, 2007; Brandtzæg and Heim, 2008; Nazir et al., 2008; Chen and Nayak, 2012). These users can be as much as 23% more influential in a social network, such as Flickr (Papagelis et al., 2011) and in the Korean SQA Naver, for example, heavy users provide 100 or more answers per week (Nam et al., 2009).

The active users in Yahoo! Answers might be opinion leaders or influential. Weimann (1994) suggests that opinion leaders are those who spread information or advice with the hopes of shaping opinions. Influential individuals were defined as a "*minority of individuals who influence an exceptional number of their peers*" (Watts and Dodds, 2007; Sakamoto et al., 2008). These users are highly relevant in understanding the diffusion of topics in the public agenda (Romero et al., 2010).

It seems that the active users in *Yahoo! Answers* have the potential of being opinion leaders. The longer the user functions as an active user, the greater are the chances he/she would set the category's agenda by asking many questions and providing numerous good answers and would become "influential" in the specific category. Hence it is not surprising that in *Yahoo! Answers* the best answers are correlated with consistent participation (Nam et al., 2009) and the highest ranking users ("*Yahoo!'s* Best Contributors") are more contributors than consumers (Shah et al., 2008).

However Yahoo! Answers' active users are not necessarily influentials (see the case of active bloggers (Agarwal et al., 2008)).

Looking at the topological characteristics of the active users, several measures were suggested for topologically identifying the core members of the network; In-degree centrality, Out-degree centrality, Closeness centrality, Betweenness centrality (Wasserman and Faust, 1994) HITS (Jurczyk and Agichtein, 2007) and PageRank (Sie et al., 2008). In addition the node's position affects collective action (Marwell et al, 1998) Kitsak et al. (2010) found that the people who are located within the core of the network are the most efficient spreaders and influences other people in adopting knowledge (Chwe, 1999). And yet, no correlation between the members' inner ranking and their position in the social network was found (Ganley and Lampe, 2009). The opinion leader's links to the neighboring vertices were more important than their overall location in the network (Valente and Davis, 1999) and the dominant members were not those with an official role in the network (Ravid and Rafaeli, 2004).

The fact that the active users have the potential of being opinion leaders (but not necessarily) and the uncertainty regarding their topological place in the network, were the main motivations to better explore their existence and importance in social networks. In the next section we will describe the network of active users, which we drawn from the overall Yahoo! Answers activities.

3. The Active Users' Network

Social networks constantly change (Tang et al., 2009). This phenomenon is particularly true in SQA where visitors seek a single piece of information (Gazan, 2011) and it seems that there is no use in referring to its average parameters or topology at all (Hill, 2009).

In this study we choose to explore the implicit-network of the active users. Based on the studies of Shi et al. (2008) and Guillaume and Latapy (2004) we build a network of active users to represent *Yahoo! Answers* activates. Following several researchers (Morgan et al., 1996; Kossinets and Watts, 2006; Viswanath et al., 2009) we try to identify the correlation between the change in the topology of the active users' network and the change in the overall activity of the entire *Yahoo! Answers*.

Following Rodrigues and Milic-Frayling (2009) and Jurczyk and Agichtein (2007) we first built a database including the 20 most active askers and 20 most active "best answerers" from each category in each month. Next, we removed all users with less than 30 contributions per month (Shi et al., 2007) and of these users we chose only the ones who were active for at least 6 months (out of 19 activity months). The active users and the content categories are the two types of nodes. The activities of the active users in the 1600 content categories are the vertices.

We explored the parameters of this simplified picture of *Yahoo! Answers,* its dynamics over time and its topology. Few studies explore the relationship between local networks' topological parameters and the whole network's topology (Barabási et al., 2002; Kossinets and Watts 2006). However, to the best of our knowledge, the idea of exploring and using the change in the topological parameters of active users' network as a substitute for exploration a huge network's dynamic is novel. This is what we propose here.

4. Hypotheses

In this section we present two hypotheses. First we'll look into the importance of active users. Is there a connection between the presence of active users in a content category and its volume of activity? In case a connection will be found, we will explore the next question. Can one use the activity and the network of the active users to depict the entire social network? Meaning, can the active users' network model a huge social network?

4.1 The importance of active users - Active users' presence and volume of activity

The SQA sites are a platform for synthetic, collaborative work with several active users who participate regularly. In H1 we inquire do content categories with active users, necessarily have overall higher volume of activity?

Intuitively, active users might attract more users to a specific content category; On the other hand the presence of a lot of users and a high volume of activity can encourage the creation of "active users" in a specific category. Since we cannot ascertain the direction of causality between the presence of major players and the category's activity, we are looking for correlation only and our first hypothesis is:

> (H1) There will be a positive correlation between the presence of major players in a content category and the overall activity of this category. In other words, there will be a significant difference in the activity level between categories with active users and those without.

4.2. The ability of active users' network – Modeling huge social network with active users' network.

In H2 we look for a greater issue: Can the network of the active users alone model the overall dynamics of activities in a huge social network?

Here we look for the correlation between the change in several topological parameters of the active users' network and the change in the activity volume of *Yahoo! Answers*. Finding such correlation gives an empirical proof to our theoretical assumption that active users' network can represent the (much bigger) entire social network.

> (H2): There is a correlation between the change in specific topological characteristics of the active users' network and the change in activity level in Yahoo! Answers.

5. Research Methodology

The data reported here consist of all the activities in Yahoo! Answers between January 1, 2009, and August 31, 2010, excluding July 2009 (which we didn't get from Yahoo!).

To choose the active users we defined and extracted 20 most active askers and 20 most active "best answerers". The overall number of records was approximately 840,000. Next we define a user as an active user if and only if: (1) The user has an average of at least one activity (asking or best answering) per day and; (2) The user was nominated in the active users records for at least 6 months of activity. We created a final list of over 1,000 active users.

Since the active users and the content categories are totally different kind of nodes, we had to build a bi-partite network. Since in real life the active users might or might not reply to each other, the only way to getting them all in one graph is to build a bi-partite network where the nodes and the categories are the nodes. The links between the nodes are the actual participation of an active user in a specific category (the same as actors and films in Kevin Bacon number). This network depicts the monthly connections between *Yahoo! Answers'* categories using the active users' activities.

In this network an active user is connected to another active user only through a mutual content category they both share in a specific month. The same goes with the categories; each two categories (or more) are connected through one or more active users who participated in both of them.

The weight of a connection between a category and an active user is the number of times an active user participated in a category. An optional weight was to use the normalized weight. We didn't use it.

In the study, we were looking basically for two correlations: (1) the correlation between the presence of major players in a content category and the overall activity of this category and (2) the correlation between the change in specific topological characteristics of the active users' network and the change in activity level in Yahoo! Answers.

The explanatory power of the correlations was evaluated by using Pearson Correlations adn the significance threshold selected was (2-tailed) Alpha <0.001.

6. Main results

In this section we present results concerning four issues: (1) The basic *Yahoo! Answers'* activity dara; (2) Data regarding the differences between categories with and without active users; (3) Data regarding the active users' network we built and (4) Data regarding the correlation between the dynamics of the active users' network and the change in the activities in Yahoo! Answers.

6.1. General data.

1. The number of categories each active user participated in is presented in Table 1.

Number of categories the active user participates in	Number of active users	Percent
1	2,997	83.9%
2	425	11.9%
3	98	2.7%
4	26	0.7%
5	15	0.4%
6 - 10	10	0.2%
Total	3,571	100.0

Table 1: The distribution of categories receiving contributions per active user in *Yahoo! Answers*

These results are in line with Chen and Nayak's, (2008, 2012) findings, where most of the answerers prefer to participate in a single (topic) category.

2.More than half (56%) of the active users were nominated as active users, for less than 9 months and only 14% were active users for most of the relevant time period. The distribution of active users' duration as is presented in Table 2. The detailed distribution is presented in table 11 in the appendix.

Duration in months	Active users' Frequency	Percent
6-9	1,999	56.0%
9-12	664	18.5%
12-15	409	11.5%
15-19	499	14.0%
Total	3,571	100.0%

Table 2: The distribution of duration of being an active user

A *significant weak negative correlation* was found between the duration of contribution of the active users as a "Best answerer" and (1) the number of the total "Best Answers" in the category (-0.18) and (2) the number of users in the category (-0.23).

These results might indicate that the period of being an active user, and hence the potential of the "Best Answers" to influence on the category, is weaken as the category size enlarges (more users and more "Best answers").

6.2. Activity's differences between categories with active users and categories without.

1. Active users' activities and parameters of size and volume of their categories.

A significant high positive correlation (0.759, N=429) was found between the number of "Best Answers" answered by the active users and (1) the number of "Best Answers" and (2) the number of questions of the category.
A significant medium positive correlation (0.6, N=429) between the number of "Best Answers" answered by the active users and the number of Answers of the category.
A significant medium positive correlation (0.52, N=139) was found between the number of questions asked by the active users and the volume of questions of the category.
A significant strong positive correlation (0.71, N=139) was found between the number of questions asked by the active users and the number of Answers of the category.

Table 3 presents the correlation's results between active users' activities and categories' size.

		Number of Best Answers	Number of Askers	Total users in category
Number of Best Answers	Pearson Correlation	1	.266**	.401**
	Sig. (2-tailed)		.002	.000
	N	429	139	429
Number of Askers	Pearson Correlation	.266**	1	-.011
	Sig. (2-tailed)	.002		.901
	N	139	139	139
Total users in category	Pearson Correlation	.401**	-.011	1
	Sig. (2-tailed)	.000	.901	
	N	429	139	429

Table 3: Correlations between active users' numbers and the number of user in the relevant categories

**. Correlation is significant at the 0.01 level (2-tailed).

2. The proportional-to-size effect of the active users.

A correlation test was performed between the relative share of the active users' activities and the category's activity volume. We tested two types of activities: supplying "Best Answers" and asking questions.

A significant weak negative correlation was found between the relative share of the "Best Answers" of the active users and The number of the total "Best

Answers” in the category (-0.21), The number of users in the category (-0.264) and The number of questions in the category (-0.209). This might indicate that active users may be a victim of their own success; the contribution, and hence the importance of the “Best Answers” to the category diminishes as the category size increases.

Next we explored and compared the size of categories with active users and without active users. In the 429 categories where active users were found: (1) The average number of users was much higher (320k users versus 2.1k); (2) The number of questions which were asked was much higher (160k versus 0.69k); (3) The total number of answers was much higher (799k average per category versus 2.8k) and (4) The total number of “Best answers” was much higher (140k versus 0.7k). The same results are apparent in the median and maximum figures.

The results suggest that there is a correlation between the category’s size and volume and the presence of active users in it. The main results are presented in Table 7 in the Appendix: Differences in Categories with and without active users.

6.3. The active users’ network topology

We built a monthly network consisting of categories and their active users as nodes and the activities as edges. The network describes how *Yahoo! Answers* is composed of many categories where the active users connect them. A path can only exist between an active user and a category, meaning that active user asked or answered in this content category in the specific month. Theoretically if each user asks and answers in one specific content category, there will be no links in this network; and in case *Yahoo! Answers’* users are active in many content categories, the network would be well connected and dense.

In the coming tables we present the analysis of the weighted, non-directional, bi-partite active users’ network through 19 activity months. We are interested in the *change* in the topological parameters of the active users’ network and its correlation with the *change* in size and volume of *Yahoo! Answers.*

The selection of each parameter and its ability to predict a change in *Yahoo! Answers’* activity is rooted in the work of Barabási et al. (2002) and to graphically analyze the network parameters we used NWB a SNA software[1]. The main parameters of the active users’ network are:

1. *The Diameter* – The shortest path between the furthest connected nodes.

2. *No of nodes* – The number of categories and their active users in each month.

3. *No of edges* – The number of active users’ activities (asking and answering) each month.

4. *The maximum weight* - The highest number of “Best Answers” or questions that were answered or asked by an active user in a category in a specific month.

5. *The average weight* - The mean number of “Best Answers” or questions that were answered or asked by an active user in a category in a specific month.

6. *The average degree* – The average number of categories in which the active user participated, OR the average number of active users each category has in a specific month.

7. *GCC* - The highest number of connected categories (even temporarily) in a specific month.

8. *Density* - The number of participated active users in a category divided by the potential of participation.

Table 4 presents the descriptive statistic of the above parameters of the active users’ network during the 19 activity months. Table 8 in the Appendix presents the network analysis for each activity month.

The degree (k) and the density are quite static during all 19 months of activity. The explanation for this phenomenon is quite simple. Since we dictated the number of users to be 20 askers and 20 best answerers as long as they were active at list 30 times per month for 6 months, we almost dictated the number of categories these users are involved with and there for the degree (through all 19 months) is between 2.07 and 2.14. Regarding the stable density, we suggested earlier that the fact that almost 84% of the active users are involved in one category only, should lead to a relatively sparse network

Since the average degree (k) and the density are quite static during all 19 months of activity, these parameters can’t explain any changes in the size or volume of *Yahoo! Answers* network. Yet, other parameters variant during the activity months and might be correlated with *Yahoo! Answers’ activity* changes. See table 8 in the appendix for full detailed table.

Parameter	Minimum	Maximum	Mean	Std. Deviation
Diameter	18	27	22.68	2.35
Nodes	4,514	5,064	4,796	149.9
Edges	4,725	5,330	5,030	171.5
Highest link Weigh - Number of interactions	1,271	2,904	1,997	419.4
Average link Weigh - Number of interactions	89	99.9	94.91	2.75
Avg Degree – Number of connecting users (k)	2.07	2.14	2.09	.016
GCC – The number of connected categories	3,714	4,316	4,042	178.82
Density	.00042	.00046	.000437	.00001

Table 4: Active users’ topological parameters - descriptive statistic of 19 months of activity

[1] (NWB Team (2006). Network Workbench Tool Indiana University, Northeastern University, and University of Michigan, http://nwb.slis.indiana.edu).

6.4 The active users' network correlation with the overall activities and the ability to depict Yahoo! Answers

We explore several parameters of *Yahoo! Answers*: The Total number of Questions, Answers, "Best Answers", users who asked and answered, users who only asked, users who only answered. Full correlation test results are presented in table 10 in the Appendix. The main results (all with alpha <0.01) are present:

A. The number of categories in the active users' network is positively correlated with: (1) The number of questions asked and (2) The number of users asking questions.

B. The number of active users in the active users' network is positively correlated with: (1) The number of questions asked and (2) The number of users asking questions

C. The contributions of the active user to the content categories in the active users' network are positively correlated with: (1) The number of answers, (2) The number of users answering, (3) The number of users both asking and answering and (4) The total number of users in *Yahoo! Answers*.

C. The average number of content categories in which the active users are active in, is positively correlated with: (1) The number of answers, (2) The number of users asking, (3) The number of users both asking and answering and (4) The total number of users in *Yahoo! Answers*.

D. The number of categories that were visited by the same users (the GCC of the users' network) is positively correlated with: (1) The total number of Questions, (2) The total number of "Best Answers" and (3) The number of askers in *Yahoo! Answers.*

E. The actual activity of the active users considering the potential of their activity (The Density of the active users' network) is negatively correlated with *all size and volume parameters* of *Yahoo! Answers.* Meaning, as more and more people are participating in *Yahoo! Answers* activities, the inner-connections between the active users are becoming sparser.

7. Discussion and Conclusions

Active users have an important role in *Yahoo! Answers'* mechanism. One can post their question to a specific category and soon enough, answers will arrive. The active users are stable, active and have a positive correlation with the activity of the whole network.

Apparently, this is the reason we found that exploring the topology of the active users' network, might depict the dynamic of all activities in *Yahoo! Answers*.

7.1 H1 findings

The data support the H1 assumption. There are several major differences between categories with active users and categories without active users. The former ones had more users, questions, answers and Best Answers. Moreover, the total number of questions per user, the total number of answers per user and the total of Best Answers per user, were dramatically different (0.32 versus 0.49, 1.3 versus 2.4 and 0.32 versus 0.42 respectively).

These results approve that there is a connection between the active users' presence and the overall activity and exploring the active users is worth a while. This finding might also suggest two casual alternative explanations:

1. Active users influence categories' parameters.

It might be sufficient for an active user to be moderately active (once a day in six months out of nineteen months) in order to influence the category size-parameters. However a higher level of contribution or longer periods as an active user does not increase the volume of activity in the category. So, it seems that even if there is an influence of the active user on the category's activity, the influence is quite limited with an upper limit.

2. Categories create their active users.

Categories with a high volume of activity and many participants create the needed social capital that enables users to deliver more inputs daily, mostly as "Best Answers", and gradually to become active users.

7.2 H2 findings

Since the active users are significant it is interesting to study their activity and network and to learn from their actions. We found a significant correlation between the dynamics of the active users' network topology and the change in *Yahoo! Answers'* activities. The active users' network reflects the activity in the *Yahoo! Answers* network, just as Barabási et al. (2002) found.

According to Barabási et al. (2002) and in contrary to the assumption of conventional models that the average distance should increase slowly as the network grows (like $O^{(\log\ n)}$), network growth over the years actually increased the average degree and GCC and decreased its diameter and CC. According to Leskovec et al. (2005) while the number of edges grew super linearly, the network density and the average distance between nodes shrunk.

Though Barabási et al.'s (2002) study explored real–life networks and not online-Q&-site-based social networks, both findings express the same dynamics. While Yahoo! answers experience an increase in number of users and number of activities, the diameter and the GCC of the active users' activity grow as well. At the same time the CC of the active users became smaller.

In our findings the only parameter that does not follow Barabási et al.'s (2002) findings is the diameter. The explanation for this is quite simple; In the active users' network, the diameter has no real meaning. Since the users act according to their changing needs, they enter and leave the SQA site as they please, there is no actual “path” between the users and hence no real diameter. Surely we can measure a diameter in active users’ network, but its value is meaningless.

For the most part our findings are applicable for researchers who wish to investigate dynamic changes such as growth or decay in huge social networks. We suggest that in order to understand the change in size and growth dynamics of *Yahoo! Answers* (20M monthly interactions), one can explore the active users’ topology (50k monthly interactions). Since there is an order of magnitude difference between the two, exploring the active users’ network should save a lot of time and efforts.

8. Study Limitations and Future Work

Since our study is based on Yahoo! Answers data only, the generalizability of the correlations should be weighed carefully. Might these findings be relevant only to this platform or to huge Q&A sites alone? Future research should examine active users in other social networks, such as Facebook or LinkedIn, to explore the external validity of the conclusions reached here.

Since we did not develop a conceptual model which includes causality to understand the direction of the mutual correlation in our hypothesis, we only tested correlations. Two regression models might be applied here; the first to present the causality and the connections between the active users’ activities and the categories’ size and volume. A second regression model might uncover the causality between the active users’ topological parameters and the size and volume parameters of the social network.

We are aware of the fact that choosing an activity threshold inevitably causes dramatically different active users’ network structures (De Choudhury et al., 2010). We defined an active user as a user who (1) acts at least on a daily basis and (2) appears at least 6 months out of 19 months of activity. These definitions generate successful results and yet a future study might consider other actions as preconditions to be considered. We believe that other insights might arise from alternative thresholds in determine active users and a sensitivity test regarding the volume of the contributions of the active player can be done.

9. Acknowledgment

We wish to thank Ricardo Baeza-Yates of Yahoo! Research Barcelona and Yoelle Maarek and her team from Yahoo! Labs Haifa for their knowledge and support. We also acknowledge with thanks the financial support of The Center for Internet Research, Univ. of Haifa, Israel.

References

L. A. Adamic, L. A., J. Zhang, E. Bakshy and M. S. Ackerman, (2008). "Knowledge sharing and yahoo answers: Everyone knows something", in Proceeding of the 17th International Conference on World Wide Web, pp. 665-674.

N. Agarwal, H. Liu, L. Tang and P. S. Yu, (2008). "Identifying the influential bloggers in a community" in Proceedings of the International Conference on Web Search and Web Data Mining, pp. 207-218

E. Agichtein, C. Castillo, D. Donato, A. Gionis and G. Mishne, (2007). "TECHNICAL REPORT YR-2007-2005”

A. Anagnostopoulos, R. Kumar and M. Mahdian, (2008). "Influence and correlation in social networks," in Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 7-15

L. Backstrom, R. Kumar, C. Marlow, J. Novak and A. Tomkins, (2008). “Preferential behavior in online groups”, Proceedings of the International Conference on Web Search and Web Data Mining, pp. 117-128

A. Barabasi, H. Jeong, H. Neda, Z. Ravasz, E. A. Schubert, and T. Vicsek, (2002). “Evolution of the social network of scientific collaborations”, Physica A: Statistical Mechanics and its Applications, 311(3-4), pp. 590-614.

M. Bouguessa, B. Dumoulin and S. Wang, (2008). "Identifying authoritative actors in question-answering forums: The case of Yahoo! Answers", in Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 866-874

P. B. Brandtzæg and J. Heim, (2008). "User loyalty and online communities: Why members of online communities are not faithful", in Proceedings of the 2nd International Conference on Intelligent Technologies for Interactive entertainment, pp. 1-10

P. Brodka, K. Musial and P. Kazienko (2009). “A performance of centrality calculation in social networks”,International Conference on Computational Aspects of Social Networks, 24-31.

L. Chen and R. Nayak, (2008). ”Expertise analysis in a question answer portal for author ranking“, in IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

L. Chen and R. Nayak (2012).“Leveraging the network information for evaluating answer quality in a collaborative question answering portal”, Social Network Analysis and Mining, 1-19.

M. De Choudhury, W. A. Mason, J. M. Hofman and D. J. Watts, (2010). “Inferring relevant social networks from interpersonal communication”, Proceedings of the 19th International Conference on World Wide Web, 301-310

M. S. Y. Chwe, (1999). "Structure and strategy in collective action”, American Journal of Sociology, pp. 128-156

D. Ganley and C. Lampe, (2009). “The ties that bind: Social network principles in online communities”, Decision Support Systems, 47(3), pp 266-274

R. Gazan, (2010). "Microcollaborations in a social Q&A community," Information Processing & Management, vol. 46, pp. 693-702

R. Gazan, (2011) "Social Q&A," Journal of the American Society for Information Science and Technology, Article first published online: 23 MAY 2011

J. L. Guillaume and M. Latapy, (2004). “Bipartite graphs as models of complex networks”, Combinatorial and Algorithmic Aspects of Networking, CAAN 2004, Banff, Alberta, Canada, August 5-7, 2004,

F. M. Harper, D. Moy and J. A. Konstan, (2009). "Facts or friends? : Distinguishing informational and conversational questions in social Q&A sites", in Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 759-768.

F. M. Harper, D. Raban, S. Rafaeli and J. A. Konstan, (2008). "Predictors of answer quality in online Q&A sites", in Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 865-874.

D. Huffaker (2010). “Dimensions of Leadership and Social Influence in Online Communities”, Human Communication Research, 36, 4, 593-617, Wiley Online Library

P. Jurczyk and E. Agichtein, (2007). “Discovering Authorities in Question Answer Communities by Using Link Analysis”, CIKM’07, November 6--8, 2007, Lisboa, Portugal.

E. Kerr, (1986). “Electronic leadership: A guide to moderating online conferences”. IEEE Transactions on Professional Communications, PC29(1), pp 12–18.

S. Kim, J. S. Oh and S. Oh, (2008). "Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective", Proceedings of the American Society for Information Science and Technology, vol. 44, pp 1-15

M. Kitsak, L, K. Gallos, S. Havlin, F. Liljeros, L, Muchnik, H. E. Stanley and H, A. Makse, (2010) “Identification of influential spreaders in complex networks”, Nature Physics, 6, 888.

J. Koh, Y. G. Kim, B. Butler, and G. W. Bock, (2007). “Encouraging participation in virtual communities”, Communications of the ACM, 50(2), pp 73

G, Kossinets and D.,J. Watts (2006). “Empirical Analysis of an Evolving Social Network, Science, Vol 31

K. R. Lakhani and E. Von Hippel, (2003) "How open source software works", Research Policy, vol. 32, pp 923-943

T. Lappas, E. Terzi, D. Gunopulos and H. Mannila, (2010) "Finding effectors in social networks," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1059-1068.

B. Li, Y. Liu and E. Agichtein, (2008) "CoCQA: Co-training over questions and answers with an application to predicting question subjectivity orientation", in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 937-946.

Y. Liu and E. Agichtein, (2008). "You've got answers: Towards personalized models for predicting success in community question answering", in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, pp. 97-100.

L. Liu, J. Tang, J. Han and S. Yang, (2012). “Learning influence from heterogeneous social networks”, Data Mining and Knowledge Discovery, 1-34, Springer

G. Marwell, P. E. Oliver and R. Prahl (1988). “Social Networks and Collective Action: “A Theory of the Critical Mass. AJS Volume 94 Number 3 pp. 502-34

A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee, (2007). “Measurement and analysis of online social networks”, Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, 42

D. L. Morgan, M. B. Neal and P. Carder,(1997). ”The stability of core and peripheral networks over time”, Social Networks, 19(1), pp. 9-25

T. Murata and S. Moriyasu, (2007). "Link prediction of social networks based on weighted proximity measures", in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 85-88.

K. K. Nam, M. S. Ackerman and L. A. Adamic, (2009). “Questions in, Knowledge-iN? A study of Naver’s question answering community”, In Proceedings of the ACM Conference on Human Factors in Computing (ACM CHI’09), pp. 779–788.

A. Nazir, S. Raza, C. N. Chuah, (2008). “Unveiling Facebook: A Measurement Study of Social Network Based Applications”, IMC’08, October 20–22, 2008, Vouliagmeni, Greece.

J. Nichols and J. H. Kang, (2012). “Asking Questions of Targeted Strangers on Social Networks”, *CSCW’12,* February 11–15, 2012, Seattle, Washington, USA

G. Palla, A. L. Barabási, and T. Vicsek, (2007). “Community dynamics in social networks”, Fluctuation and Noise Letters, 7(3), pp 273-287

S. A. Paul, L. Hong and E. H. Chi, (2011). “Is Twitter a Good Place for Asking Questions? A Characterization Study”, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011

M. Papagelis, V. Murdock and R. van Zwol, (2011). "Individual behavior and social influence in online social systems," Work, vol. 17, pp. 31

D. Raban and F. Harper, (2008). “Motivations for answering questions online”, New Media and Innovative Technologies

S. Rafaeli and R. J. LaRose, (1993). ”Electronic bulletin boards and public goods explanations of collaborative mass media“, Communication Research, vol. 20, pp. 277

G. Ravid and S. Rafaeli. (2004). “Asynchronous discussion groups as small world and scale free networks”, First Monday, 9(9-6)

A. Rechavi and S. Rafaeli, (2012) "Knowledge and Social Networks in *Yahoo! Answers*", HICSS 45, Hawaii, 2012.

M. Richardson and P. Domingos, (2002) "Mining knowledge-sharing sites for viral marketing," in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 61-70.

M. E., Rodrigues and N., Milic-Frayling, (2009). “Socializing or knowledge sharing?: characterizing social intent in community question answering” Proceedings of the 18th ACM conference on Information and knowledge management pp. 1127-1136,

D. M. Romero, W. Galuba, S. Asur and B. A. Huberman (2010).“Influence and passivity in social media”, arXiv:1008.1253v1 [cs.CY] 6 Aug.

C. Shah, L. S. Oh and S. Oh (2008). “Exploring characteristics and effects of user participation in online social Q&A sites”, First Monday, 13(9)

X. Shi, L. A. Adamic and M. J. Strauss (2007). “Networks of strong ties”, Physica A: Statistical Mechanics and its Applications, 378(1), pp. 33-47

X. Shi, M. Bonner, L. A. Adamic and A.C Gilbert (2008). “The very small world of the well-connected”, Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia, pp. 61-70

V. Soroka and S. Rafaeli, (2006). "Invisible participants: How cultural capital relates to lurking behavior", in Proceedings of the 15th International Conference on World Wide Web, pp. 172

M. Trusov, A. V. Bodapati and R. E. Bucklin, (2010). “Determining influential users in internet social networks”, Journal of Marketing Research, 47(4), pp. 643-658

T. C. Turner, M. A. Smith, D. Fisher and H. T. Welser,(2005). "Picturing Usenet: Mapping computer-mediated collective action," Journal of Computer-Mediated Communication, vol. 10, pp. 7.

T. W. Valente and R. L. Davis (1999) “Accelerating the diffusion of innovations using opinion leaders”, The Annals of the American Academy of Political and Social Science, 566(1), 55.

B. Viswanath, A. Mislove, M. Cha and K. P. Gummadi, (2009). ”On the evolution of user interaction in Facebook“, in Proceedings of the 2nd ACM Workshop on Online Social Networks, pp. 37-42

S. Wasserman, K. Faust, (1994). “Social network analysis: Methods and applications”, New York: Cambridge University Press, 1994

D. J. Watts and P. S. Dodds, (2007). "Influentials, networks, and public opinion formation”, Journal of Consumer Research, vol. 34, pp. 441-458

G. Weimann, (1994). “The influentials: People who influence people”, Albany, NY: State University of New York Press, 1994.

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

J. Yang, X. Wei, M. S. Ackerman and L. A. Adamic, (2010). “Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities”, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

D. Zinoviev (2008). “Topology and geometry of online social networks”, Proc. 12th World Multi-Conference on Systemics, Cybernetics and Informatics VI pp. 138-143

10. Appendix

Table 5: Correlations between active users' activity and the relevant categories volume of activity

		Mean contribution as Best answerer	Mean contribution as asker	Total Answers in the category	Total Best Answers in the category	Total Questions in the category
Mean contribution as Best Answer	Pearson Correlation	1	.578**	.601**	.759**	.759**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	429	139	429	429	429
Mean contribution as asker	Pearson Correlation	.578**	1	.710**	.585**	.527**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	139	139	139	139	139
Total Answers in the category	Pearson Correlation	.601**	.710**	1	.900**	.855**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	429	139	429	429	429
Total Best Answers in the category	Pearson Correlation	.759**	.585**	.900**	1	.994**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	429	139	429	429	429
Total Questions in the category	Pearson Correlation	.759**	.527**	.855**	.994**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	429	139	429	429	429

Table 5: Correlations between active users' activity and the relevant categories volume of activity

Table 6: Correlations between active users' share and parameters in the relevant categories

		Number of users	Share of best answerers	Number of Best Answers	Number of answers	Number of questions
Number of users in category	Pearson Correlation	1	-.264**	.916**	.710**	.938**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	429	429	429	429	429
Share of best answerers in category	Pearson Correlation	-.264**	1	-.210**	-.133**	-.209**
	Sig. (2-tailed)	.000		.000	.006	.000
	N	429	429	429	429	429
Number of Best Answers	Pearson Correlation	.916**	-.210**	1	.900**	.994**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	429	429	429	429	429
Number of answers in category	Pearson Correlation	.710**	-.133**	.900**	1	.855**
	Sig. (2-tailed)	.000	.006	.000		.000
	N	429	429	429	429	429
Number of questions in category	Pearson Correlation	.938**	-.209**	.994**	.855**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	429	429	429	429	429

Table 6: Correlations between active users' share and parameters in the relevant categories

Table 7: Differences in Categories with and without active users

Categories with and without active users		Mean number of users	Mean number of Questions	Mean number of Answers	Mean number of BA
With active user	Mean	326,369	160,026	799,279	140,238
	N	429	429	429	429
	Std. Deviation	544,514	342,814	2,748,595	297,814
	Median	117,310	47,504	185,969	43,102
	Maximum	5,217,921	3,659,713	44,841,364	3,575,315
Without active user	Mean	2,119	699	2,808	711
	N	1,221	1,222	1,217	1,217
	Std. Deviation	6,650	1,811	8,951	1,753
	Median	454	127	416	132
	Maximum	149,672	28,057	203,634	25,681

Table 7: Differences in Categories with and without active users

Table 8: active users' Network parameters in 19 activity months

Month Year	Diameter	Nodes	Edges	Max Weight	Mean Weight	Average Degree	WCC count	(No. of Cat in the) GCC	Density
01-2009	22	4,685	4,895	1,271	89	2.09	116	4,011	0.00045
02-2009	22	4,514	4,725	2,349	89	2.1	117	3,876	0.00046
03-2009	22	4,923	5,180	1,538	94	2.1	134	4,218	0.00043
04-2009	22	4,748	5,023	1,523	93.6	2.11	123	4,119	0.00045
05-2009	20	4,825	5,165	2,108	95.7	2.14	108	4,291	0.00044
06-2009	25	4,933	5,192	1,774	93.4	2.11	122	4,204	0.00043
08-2009	26	5,064	5,330	2,196	94.5	2.11	113	4,266	0.00042
09-2009	23	4,898	5,157	2,471	95.6	2.11	121	4,055	0.00043
10-2009	24	4,888	5,158	1,815	97.5	2.11	118	4,316	0.00043
11-2009	22	4,779	5,000	1,760	93	2.09	130	3,981	0.00044
12-2009	18	4,543	4,769	1,717	94.9	2.1	138	3,808	0.00046
01-2010	24	4,739	4,948	2,406	95.0	2.09	147	3,774	0.00044
02-2010	20	4,746	4,949	2,155	93.5	2.08	142	3,714	0.00044
03-2010	23	5,050	5,301	1,779	96.8	2.1	136	4,160	0.00042
04-2010	27	4,738	4,984	1,615	97.6	2.08	145	3,984	0.00044
05-2010	26	4,689	4,869	1,776	96.6	2.07	151	3,896	0.00044
06-2010	21	4,655	4,838	2,444	95.9	2.08	139	3,953	0.00045
07-2010	20	4,919	5,109	2,904	99.9	2.08	142	4,035	0.00042
08-2010	24	4,792	4,987	2,356	97.9	2.08	125	4,152	0.00043

Table 8: Active Users Network. Parameters in 19 activity months.

Table 9: Active Users Network - Correlations of the Parameters in 19 activity months

		Diameter	No of Nodes	No of Edges	Max Weight	Mean Weight	Average Degree	Nodes in the GCC:	Density
Diameter	Pearson Correlation Sig. (2-tailed)	1	.334 .162	.299 .213	-.164 .501	.158 .518	-.151 .536	.271 .262	-.389 .100
No of Nodes	Pearson Correlation Sig. (2-tailed)	.334 .162	1	.977 .000	.075 .761	.401 .089	.313 .193	.698 .001	-.939 .000
No of Edges	Pearson Correlation Sig. (2-tailed)	.299 .213	.977 .000	1	-.001 .997	.350 .141	.491 .033	.771 .000	-.867 .000
Max Weight	Pearson Correlation Sig. (2-tailed)	-.164 .501	.075 .761	-.001 .997	1	.386 .103	-.150 .540	-.163 .505	-.254 .295
Mean Weight	Pearson Correlation Sig. (2-tailed)	.158 .518	.401 .089	.350 .141	.386 .103	1	-.183 .454	.232 .340	-.572 .011
Average Degree	Pearson Correlation Sig. (2-tailed)	-.151 .536	.313 .193	.491 .033	-.150 .540	-.183 .454	1	.623 .004	-.062 .801
Nodes in the Giant connected component	Pearson Correlation Sig. (2-tailed)	.271 .262	.698 .001	.771 .000	-.163 .505	.232 .340	.623 .004	1	-.568 .011
Density	Pearson Correlation Sig. (2-tailed)	-.389 .100	-.939 .000	-.867 .000	-.254 .295	-.572 .011	-.062 .801	-.568 .011	1

Table 9: Active Users Network. Correlations of the Parameters in 19 activity months

Table 10: The Correlation between Active Users Network Topology and Yahoo! Answers activity's parameters

		Q total	Answers BA	Answers Total	Users both asking and answering	Users_Only Asking	Users Only Answering	Users Total
Diameter	Pearson Correlation	.233	.032	-.046	-.018	.196	-.021	.002
	Sig. (2-tailed)	.336	.895	.852	.940	.422	.932	.995
Nodes	Pearson Correlation	.609	.431	.191	.252	.601	.139	.200
	Sig. (2-tailed)	.006	.066	.433	.297	.006	.570	.412
Edges	Pearson Correlation	.660	.502	.284	.340	.674	.229	.289
	Sig. (2-tailed)	.002	.028	.240	.154	.002	.346	.230
Weight Max	Pearson Correlation	-.246	-.327	-.387	-.270	-.292	-.390	-.378
	Sig. (2-tailed)	.310	.171	.102	.264	.225	.098	.110
Weight Mean	Pearson Correlation	-.235	-.426	-.656	-.587	-.304	-.679	-.646
	Sig. (2-tailed)	.333	.069	.002	.008	.206	.001	.003
Average Degree	Pearson Correlation	.566	.566	.593	.623	.670	.561	.589
	Sig. (2-tailed)	.012	.011	.007	.004	.002	.012	.008
GCC	Pearson Correlation	.753	.631	.451	.474	.718	.414	.460
	Sig. (2-tailed)	.000	.004	.052	.040	.001	.078	.048
Density	Pearson Correlation	-.682	-.684	-.643	-.631	-.708	-.619	-.642
	Sig. (2-tailed)	.001	.001	.003	.004	.001	.005	.003

Table 10: The Correlation between active users' Topological and Yahoo! Answers activity's parameters. Alpha < 0.001

Table 11: The Distribution of being an active user (in months) in all categories in Yahoo! Answers

Time period in Months		Frequency	Percent
How many months the active users were "active users" in a row	6.00	767	21.5
	7.00	497	13.9
	8.00	365	10.2
	9.00	273	7.6
	10.00	228	6.4
	11.00	205	5.7
	12.00	140	3.9
	13.00	153	4.3
	14.00	113	3.2
	15.00	87	2.4
	16.00	91	2.5
	17.00	79	2.2
	18.00	75	2.1
	19.00	127	3.6
	Total	3,571	100.0

Table 11: The Distribution of the time period of being an active user in all categories in Yahoo! Answers