Cultivating online: an analysis on an agricultural Q&A community

Li Xiang[1], Gu Liyang[1], Chen Xin[1], Liu Lei[2], Jia Lu[1*]

(1. *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*;
2. *Shandong Laodao Network Technology Co. LTD., Shandong 261000, China*)

Abstract: The "wisdom of the crowd" phenomenon has revolutionized content generation, discovery and curation in the Internet age. Online question & answer (Q&A) communities are nowadays enlightening over a billion people with crowdsourced knowledges. While previous analyses often focus on general Q&A sites like Quora and Yahoo answers, in this article, we conduct an in-depth analysis on an online Q&A site named "Nong Guan Jia" (meaning farm butler) that is exclusive for agricultural knowledges. Based on datasets containing over seven thousand questions and over 66 thousand answers, and detailed user information of over two thousand experts and over four thousand farmers, we characterize its knowledge repository and user activities, demonstrate positive reinforcement between user activity and popularity, propose a graph model that reveals user relationships and high-level structures, and successfully apply our findings to build machine-learnt classifiers to identify potential active and popular experts. Our analyses provide valuable information for maintaining the community prosperity and therefore assist the development of agricultural knowledge sharing.

Keywords: online Q&A community, user behavior, graph model, prediction

Citation: Li, X., L. Y. Gu, X. Chen, L. Liu, and L. Jia. 2017. Cultivating online: an analysis on an agricultural Q&A community. International Agricultural Engineering Journal, 26(3): 283–291.

1 Introduction

In China, with the development of agricultural industry and information technology, the number of online e-commerce platform about agriculture is exploding, reaching more than 30,000 till the present moment. Meanwhile, online question & answer (Q&A) communities, embodying the "wisdom of the crowd", are nowadays a major Internet phenomenon that educate over a billion users. While previous studies often focus on general Q&A sites, in this article, we choose Farm-Butler, an online Q&A site that is exclusive for agricultural knowledge as our research object.

Given the user scale, dynamics, and decentralization of the contents provided by individual users and the gradually shifting interests of the users, two fundamental questions for maintaining and growing such Q&A sites are that, at the questioner's perspective, how to improve the response rate of their questions so as to achieve smooth user questioning experiences, and that, at the replier's perspective, how to collect more rewards that will encourage them to maintain their activity level. In this paper, we seek to combine these two tasks so as to promote the community prosperity and to assist the development of agricultural knowledge sharing.

Our analysis of Farm-Butler mainly consists of three parts. First, we reveal, quantitatively, the scale and the characteristics of Farm-Butler by examining the question repository and user activities, and demonstrate a positive reinforcement between user activity and popularity. Then, we propose a graph model based on the Q&A activities, which captures both the direct user relationships and the higher-order social structures. Finally, applying our findings, we develop two machine-learned classifiers that can successfully identify experts who will reply a lot of questions and who will receive a large amount of donations, respectively.

Received date: 2017-06-19 **Accepted date:** 2017-08-17

* **Corresponding author: JIA Lu,** Ph.D., Associate Professor of College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China. Email: ljia@cau.edu.cn.

We summarize our contributions as follows:

- We collect four datasets that contain the complete view of Farm-Butler, with detailed statistics for 7,870 questions, 66,558 answers, 4,378 farmers, and 2,774 experts (Section 2).
- We provide a characterization on Farm-Butler. Our analyses include (i) the repository scale, (ii) the statistical properties of the question popularity, (iii) the questioning activity and the collected attention of the farmers, (iv) the reply activity and expert popularity of the experts, and (v) the user location distributions (Section 3).
- We propose a directed graph model to analyze the user relationships. The model contains in total 5,318 users (Section 4).
- We build machine-learned classifiers to predict with high accuracies the experts that will be active and gain great popularity (Section 5).

2 Farm-Butler dataset

2.1 An overview of Farm-Butler

Farm-Butler is an online Q&A site like Quora and Yahoo Answers but is exclusive for agricultural knowledge. Similar to general Q&A sites, the platform allows the users to raise questions that other users can answer directly. All users in Farm-Butler can browse the information and give the thumbs up or the thumbs down to the answers they find useful or wrong. In addition, Farm-Butler provides a feature of donation wherein users can donate to other users, in real money, to show the support for their efforts.

Compared with the general Q&A sites, it's more difficult for a questioner in Farm-Butler to act as a replier in the meantime, due to the exclusiveness and limitation of agricultural knowledge. For this reason, Farm-Butler has manually labeled the user status as either expert or farmer, based on the following rules:

(i) Farmers: the users that have raised at least one question,

(ii) Experts: the users that have replied at least one question.

2.2 Dataset

For our analysis, we have collected four datasets named the question dataset, the answer dataset, the farmer dataset and the expert dataset, respectively. The former two datasets represent a sample of the whole Q&A repository. They were collected on the 1st and the 15th of each month for the year of 2016. The latter two datasets include the complete user information since they join the community.

More specifically, the question dataset contains, for each question, the question identification (ID), the time when it was raised, the province and the city of the questioner, and when and who have replied. The answer dataset contains, for each answer, the answer ID, the question id of the associated question, the replier ID, and the reply time.

The farmer dataset contains, for each farmer, the farmer ID, the farmer's name, the province and the city of the farmer, the number of questions he has raised, and the number of replies received by all his questions. The expert dataset contains similar information, and in addition, for each expert, the number of ups, the number of downs, and the number of donations. The basic statistics of our datasets are introduced in Table 1.

Table 1 Farm-Butler scale

Type	Number	Type	Time dimension
#question	7,870	question time	1st and 15th of Jan 2016 – Dec 2016
#answer	66,558	reply time	1st and 15th of Jan 2016 – Dec 2016
#farmer	4,378	time	Jan 2016 – Dec 2016
#expert	2,774	time	Jan 2016 – Dec 2016

3 Farm-Butler characteristics

In this section, we first introduce the scale of Farm-Butler. Then we provide a characterization on its question repository and the user activities.

3.1 Farm-Butler scale

Table 1 introduces the scale of Farm-Butler derived from our datasets. During our observation period of 36 days extending one year, 7,870 questions were raised and 66,558 answers were made, achieving an average of 218 questions per day and 8 answers per question. Altogether, 4,378 farmers and 2,774 experts are involved, resulting an average of 24 answers per expert.

3.2 Question characteristics

In this section, we first discuss the temporal and spatial distribution of the questions. Then, we study the replies of the questions to examining their popularity.

3.2.1 Question injection

Figure 1 shows the number of questions injected on the 1st and 15th of each month in 2016. We find that Farm-Butler is growing dramatically from February to June, and users are most active during the summer time, i.e., from May to August. We believe this is the harvest time for most farmers and therefore they urgently need helps and advices from the experts.

To take a closer look, Figure 2 depicts the number of questions injected at different time of the day (in hour). We find that questions are mostly raised during the daytime, without a specific preference for the exact hour. This phenomenon is different from the patterns that have been observed in other online communities, such as YouTube and Twitch (Figueiredo et al., 2011; Jia et al., 2016). One possible reason is that most farmers do not follow a nine-to-five work as people in the urban areas do and therefore they have plenty of time to go online during the daytime.

3.2.2 Question location

Question location directly measures the usage of Farm-Butler in different areas. Figure 3 shows the number of questions from different provinces and autonomous regions in China. We find that Shandong Province is the top province in terms of number of questions, accounting for 14% of the total question repository. The other top nine provinces or autonomous regions account for more than 56% of the questions, while the remaining locations account for 30%.

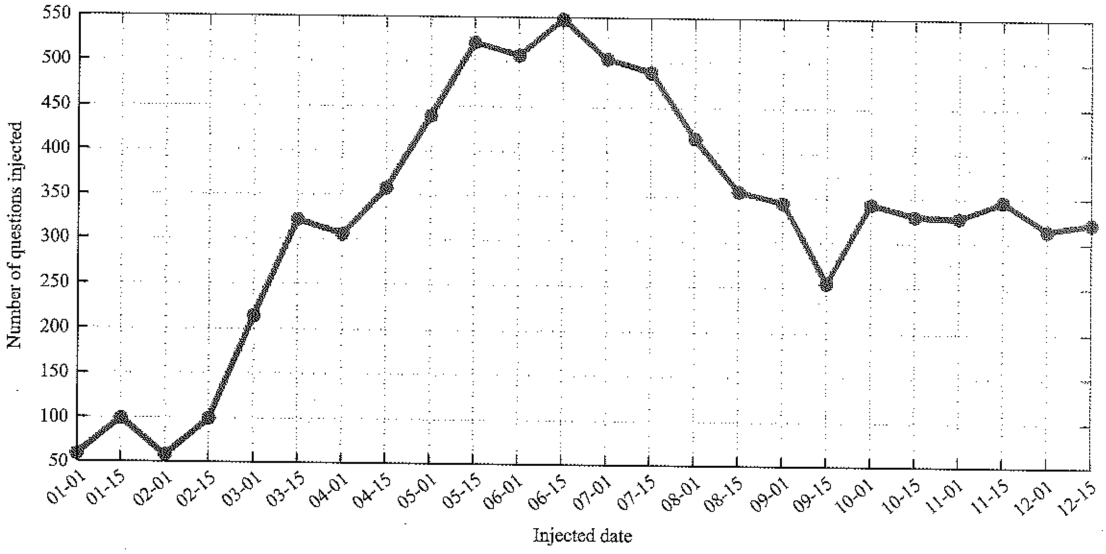

Figure 1 Number of questions injected over time

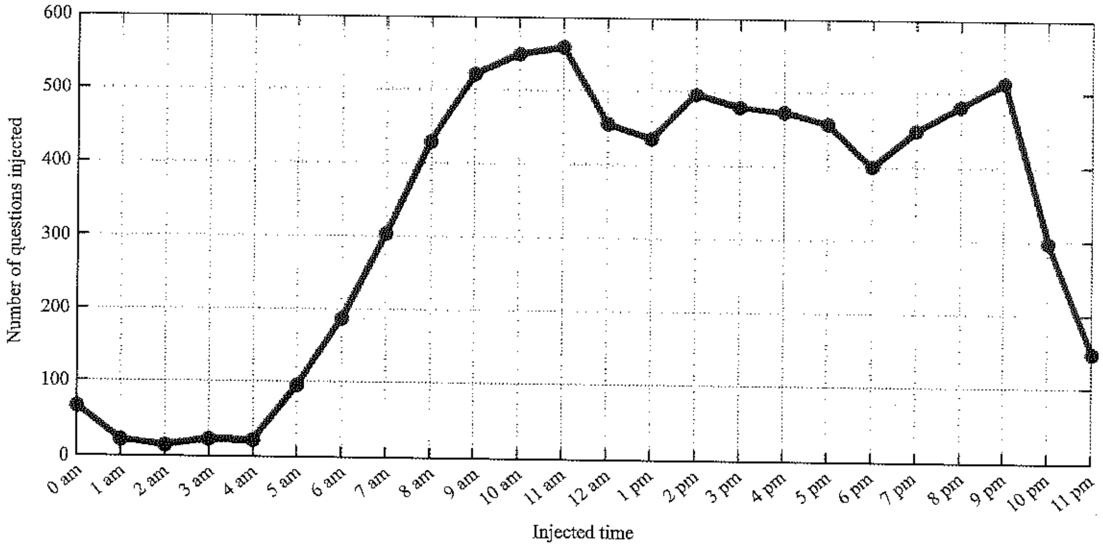

Figure 2 Number of questions injected at different time of a day

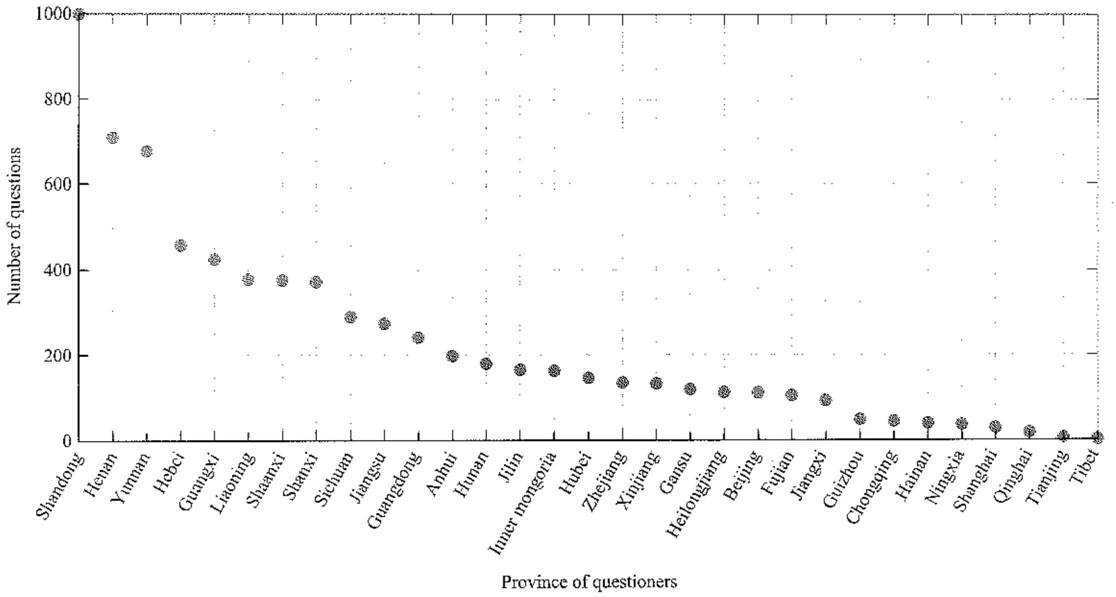

Figure 3 Geographic distribution of the questions

3.2.3 Question popularity

In any online Q&A site, question popularity provides important knowledge for the activity level of the experts, the potential workload for maintaining the site, and the community prosperity. Here, we measure the question popularity in the terms of the number of expert's replies to the question.

It is shown in in Figure 4 the complementary cumulative distribution function (CCDF) of the number of replies collected by each question. The question popularity is highly skewed, with a small number of questions attracting a large number of replies. When it is plotted on a log-log scale, we observe a curve that is close to a straight line with a negative slope, indicating that the number of replies can be well described as a power-law distribution.

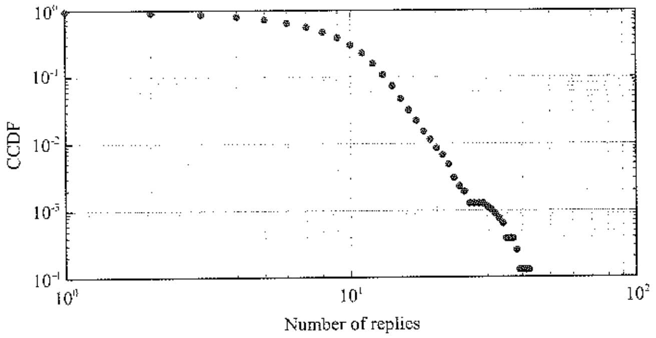

Figure 4 CCDF of the number of replies of the questions

3.3 User activities

In this section, we first analyze the activity and the popularity of the farmers and experts. Then, we study the location distribution of the users.

3.3.1 Questioning activity and the collected attention

In Farm-Butler, 4,378 users have raised at least one question and we name them farmers. The farmers on average have questioned 18.18 problems and have collected 247,789 replies, resulting in an average of 56.60 replies per question. The detailed statistics are shown in Table 2.

Table 2 Basic statistics of the farmers

	Number of questions	Number of replies
Minimum	1	0
1st quartile	4	1
Median	9	5
Mean	18.18	56.60
3rd quartile	21	15
Maximum	2,215	66,565

Questioning activity. Figure 5 plots the number of questions raised by the farmers, with the farmers ranked in the decreasing order by the number of questions. Surprisingly, the number of questions does not follow a Zipf distribution, indicating that the activity level in Farm-Butler is not as highly skewed as often observed in other online communities (Ding et al., 2011). We conjecture that the high response rate of questions in Farm-Butler promotes users to raise more questions and hence reduce the disparity. Among these farmers, the most active 20% of the farmers contribute 62.64% of the questions.

Collected attention. We measure the attention that a farmer collected by the number of replies received by all the questions he raised. Figure 6 plots the CCDF of the number of replies received by the farmers. We find that 66.95% farmers have received fewer than 10 answers while around 1% farmers have received more than 1,000 answers for their questions. The disparities in the attentions they get are highly likely due to the differences in the number of questions they raised.

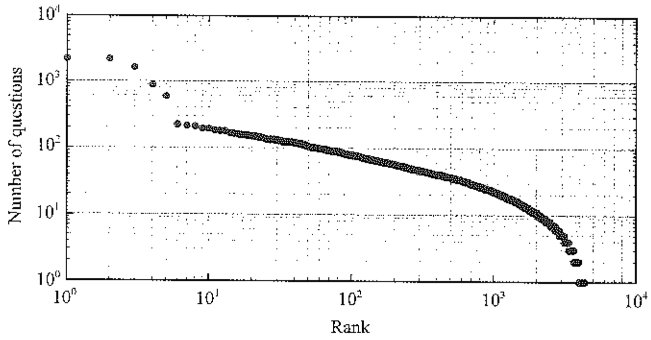

Figure 5 Number of questions of the farmers, ordered in the decreasing order

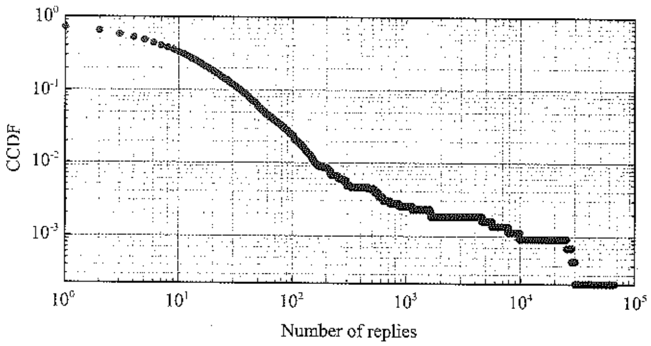

Figure 6 CCDF of the number of the replies collected by the farmers

3.3.2 Reply activity and expert popularity

In total, 2,774 users have replied at least one question and are labeled as experts by the community. We use the number of replies of the experts to measure their activity level. On average, each expert has replied 431 questions.

Reply activity. As shown in Figure 7, for experts in Farm-Butler, their levels are highly skewed: while 88.79% of experts have replied fewer than 100 questions, 2.81% of experts have replied more than 1,000 questions.

Expert popularity. Farm-Butler designed two features for the users to evaluate the answers made by the experts. They can give the thumbs ups or downs to the replies they find of high quality and they can donate to the experts in real money. In total, 1,369 experts have received at least one up, with a maximum number of ups, downs, and donations of 17,426, 2,542 and 27,146, respectively. To take a closer look, Figure 8 shows the number of ups, downs, donations, received by each expert, with the expert ranked in the decreasing order. We find that the expert popularity is high skewed.

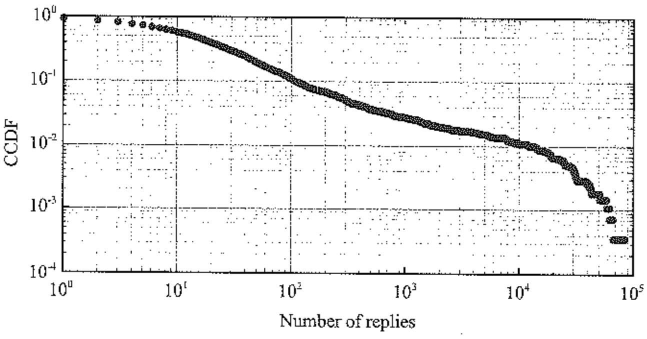

Figure 7 CCDF of the number of replies of the experts

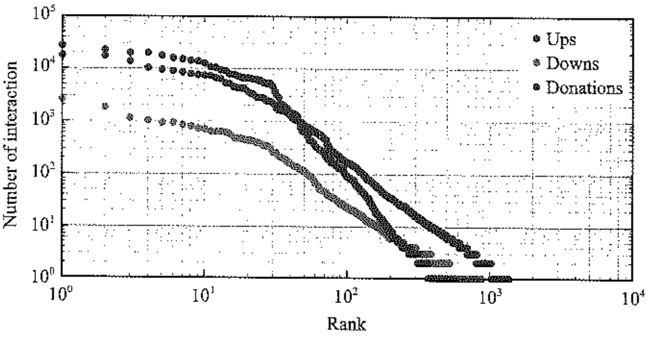

Figure 8 The number of ups, the number of downs, and the number of donations received by each expert, ranked in the decreasing order

Table 3 shows the spearman ranking correlation coefficients (SRCCs) and Pearson correlation coefficients (PCCs) between the number of ups, the number of downs, and the number of donations. We observe high correlations between any two of the three measures, and the SRCC between the number of downs and the number of donations reaches 0.9472. The high correlation between these features will later be leveraged for our prediction tasks.

Table 3 Correlations between number of ups, n_u, number of downs, n_d, and number of donations, n_a

	n_u vs. n_d	n_u vs. n_a	n_d vs. n_a
SRCC	0.8734	0.8401	0.9472
PCC	0.6602	0.5461	0.5102

3.3.3 User locations

Figure 9 and Figure 10 show the location distribution of farmers and experts, respectively. Similar to the location distribution of the questions (as show in Figure 3), the top ten areas are identical in terms of the number of questions, farmers and experts, indicating the using level of farmers and experts in Farm-Butler is similar.

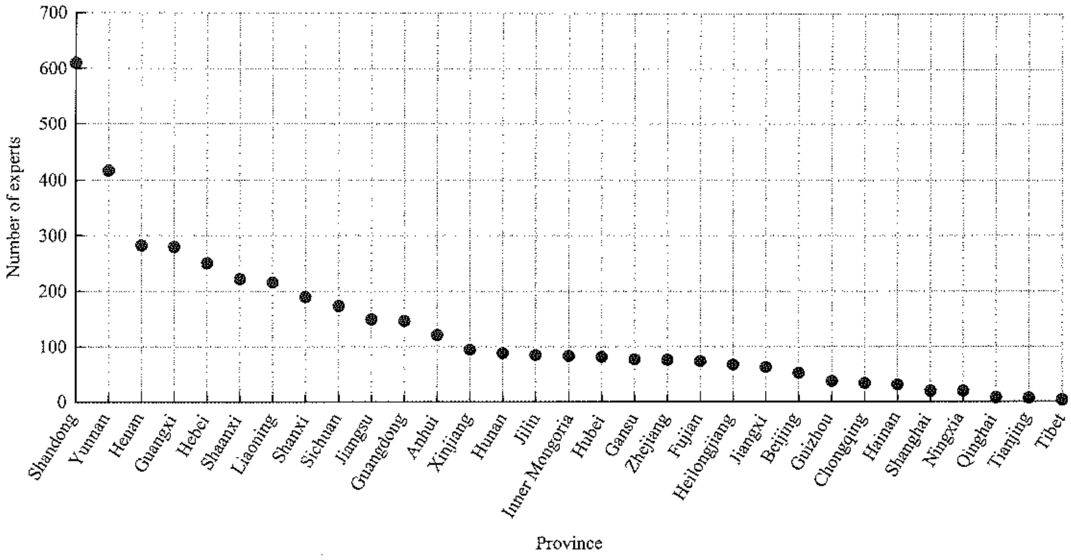

Figure 9 Geographic distribution of the farmers

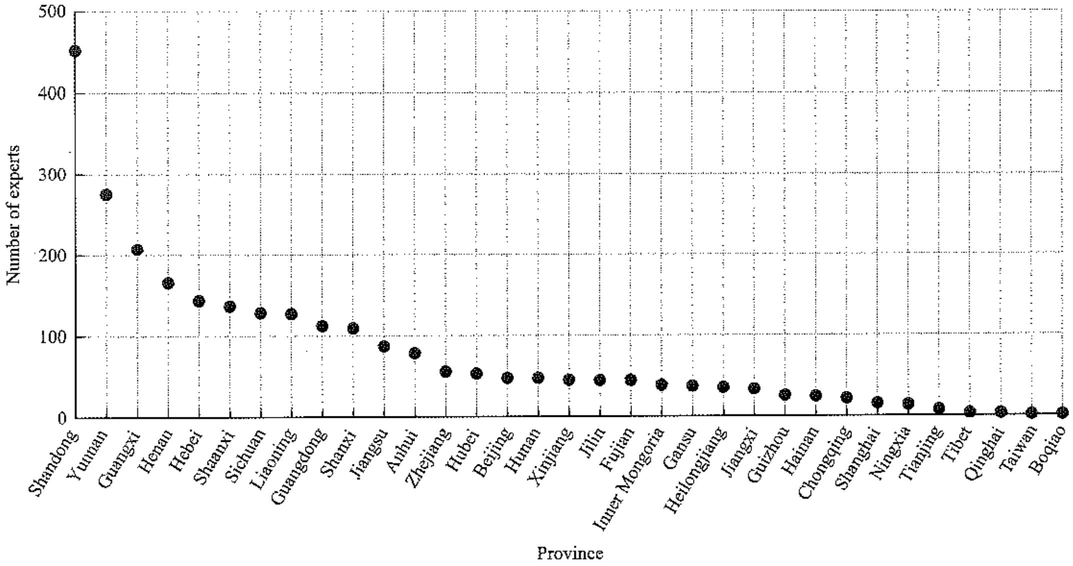

Figure 10 Geographic distribution of the experts

4 Graph model

To analyze user relationships in Farm-Butler, in this section we propose a graph model, named reply graph, based on expert's reply activities. In the reply graph, a vertex represents a user, and an edge directed from one vertex to another represents a user replies another user's questions. The edge weight indicates the number of questions in which the two users are involved. This graph model captures both the direct user relationships and higher-order social structures.

4.1 Edge weight distribution

Figure 11 shows the cumulative distribution function (CDF) of the edge weight of the reply graph. We find that, 82.68% of the edges have a weight of 1, indicating that most user pairs have only encountered once. Nevertheless, 305 edges have a weight of more than 5, and 158 edges have a weight of more than 10, showing that a modest number of user pairs have engaged repeatedly.

4.2 Degree distribution

Since the reply graph is directed, vertices have both in-degree (the number of respondents to the questions raised by the user) and out-degree (the number of cases wherein the users act as a respondent).

Figure 12 shows the CCDF of the in-degree and the out-degree for the experts and the farmers, respectively.

All the farms have out-degree of zero, as they have never answered others' questions. When plotted on a log-log scale, for the in-degree of the farmers and experts, we see a heavy tail resembling a power-law distribution. Surprisingly, there is very little variation between the in-degree distribution of farmers and experts, indicating that experts are also accustomed to raise questions in Farm-Butler. It is assumed that part of the experts are also farmers - they learn from the community and at the same time help others when they can.

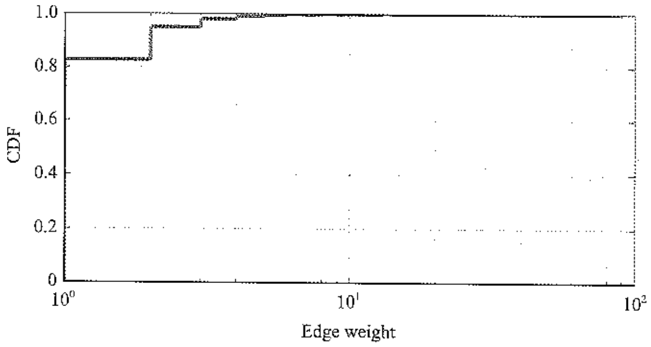

Figure 11 CDF of edge weight

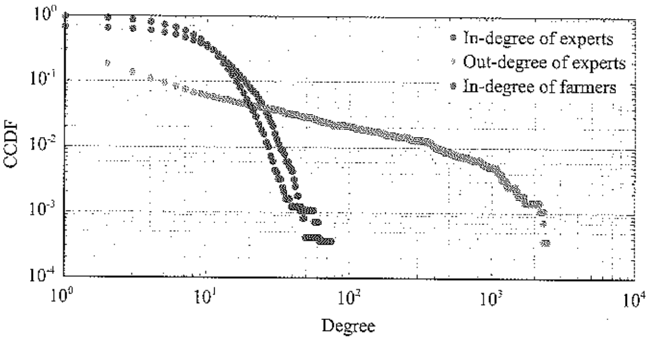

Figure 12 CCDF of in-degree and out-degree distributions

4.3 Clustering coefficient

The clustering coefficient (Watts and Strogatz, 1998) in social networks measures the fraction of users whose friends are themselves friends (Myers et al., 2014). Here, we examine the clustering coefficient of vertices in the Farm-Butler reply graph. Figure 13 plots the mean of the clustering coefficient against the vertex degree. We find that for small degrees (1-50), the average clustering coefficient stays stale, with minor fluctuations. When node degree further increases, the average clustering coefficient drops dramatically (notice the log scale for the axis), showing that with more "friends", it is more difficult to keep them close. Similar phenomenon has been observed in many other online social networks (Myers et al., 2014).

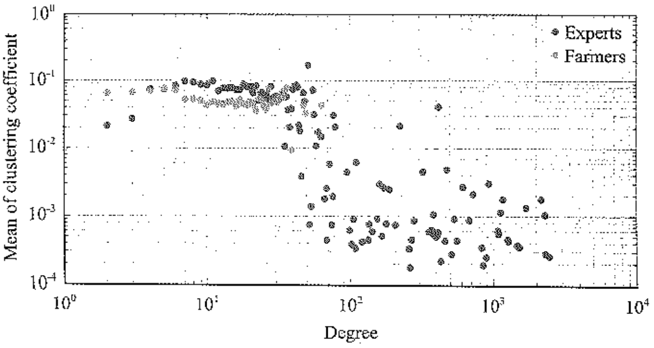

Figure 13 Clustering coefficient versus node degree

5 Predicting priority experts

Having gained several valuable insights on the characteristics of Farm-Butler and Farm-Butler users, we are now in a position to apply these findings by building machine-learned classifiers to predict experts that will be top active and gain great popularity. Successfully identifying these experts provides valuable information for maintaining and for boosting the community prosperity.

Classification tasks and methodology. More specifically, we have two classification tasks, one is to predict whether an expert will be one of the top experts in terms of the number of replies he responds, the other is to predict whether an expert will be one of the top experts in terms of the number of donations it received.

To this end, we have collected information on all the 2,774 experts that have replied questions during our observation period. At the end of this observation, we find that 311 experts have replied more than 100 questions, being the top 11% active experts, and 96 experts have received more than 100 donations, being the top 3.5% active experts. For the two classification tasks, we prepare two datasets, labeling the top 11% and top 3.5% experts as positive examples and the rest as negative examples, respectively.

Classification algorithm. We experimented with two classification algorithms-support vector machines, and random forests, and found the latter to work best. Hence all results reported here were obtained using random forests (Breiman, 2001). For each experiment, we run 5-fold cross validation and report the Area Under the receiver operating characteristic (ROC) Curve (AUC). We use balanced training and test sets containing equal numbers of positive and negative example, so random guesting results in an AUC of 50%.

Features. Based on previous analysis, we extract two groups of features including the user characteristic (u), and the reply graph properties (g). All these features have been extensively studied in Sections 3 and 4, and are summarized in Table 4. For the classification task of predicting the expert activity, we remove the number of replies in the user characteristic feature group. Similarly, for the classification task of predicting the expert popularity, we remove the number of donations in the user characteristic feature group.

Table 4 Classification features

Feature group	Description
User characteristic (u)	Number of replies, ups, downs, and donations
Reply graph properties (g)	Degree (in and out), clustering coefficient, and PageRank score of the experts

Results. Results for the two classification tasks are shown in Table 5 and Table 6, respectively. In order to understand which features are important for the prediction, we have progressively increased the group of features used in the classifier. We have a number of interesting findings as follows.

Table 5 Classification results of predicting replies

Features	AUC
u	90.83%
g	86.80%
u+g	93.37%

Table 6 Classification results of predicting donations

Features	AUC
u	94.44%
g	92.78%
u+g	95.00%

First, for the two tasks, using only the user characteristics achieves an AUC of 90.83% and 94.44% (remarkably better than random guessing). It indicates that the correlation between user characteristics is high, confirming our observation of high correlation between the number of ups, the number of downs, and the number of donations as introduced in section 3.3.3.

Secondly, the AUC of the two classification tasks reach 86.80% and 92.78% using only the reply graph properties. This result shows that the reply graph model we proposed provide valuable information for the two prediction tasks.

Thirdly, the best performance is achieved when combining all feature groups: the classifiers achieve an AUC of 93.97% and an AUC of 95.00% respectively for the two prediction tasks.

6 Related work

Related work is summarized within each research topic our work covers as follows.

Q&A in social networks. It refers to people asking questions on social networking sites. Morris et al. (2010) have explored how users leverage general-purpose online social networks for information seeking. Paul et al. (2011) conduct a study of question asking and answering behavior on Twitter and found that most popular question types were rhetorical and factual. Researchers evaluate the role of tie strength in question answers, finding that stronger ties (close friends) provide a subtle increase in information gain (Panovich et al., 2012). Users also can ask their friends questions by updating status in Facebook (Morris et al., 2010). These studies analyze users' behavior and the social network characteristic.

Community based Q&A. Researchers have studied community Q&A sites like Yahoo! Answers (Adamic et al., 2008), Live Q&A (Rodrigues et al., 2008) and MSN QnA (Hseih et al., 2009). For Live Q&A and Yahoo! Answers, Rodrigues et al. (2008) provide an in-depth analysis of the question labeling practices, finding that community tagging is related to higher levels of social interactions amongst users. Some studies explore the use of machine learning techniques to automatically classify questions as conversational or informational (Harper et al., 2009). Others estimate the quality of user's questions and answers (Paul et al., 2011; Shah and Pomerantz, 2010). In addition, some studies aim to develop algorithms to identify users with high capacity. For example, Pal et al. (2012) analyze the changes in experts' behavioral patterns over time, and using unsupervised machine learning methods to distinguish experts from one another.

Different from these work, we choose Farm-Butler, an online Q&A site that is exclusive for agricultural knowledge as our research object, focus on the experts' activity and popularity by investigating the question repository and the user activities.

7 Conclusion

In this paper, we conducted an analysis on an online Q&A site named Farm-Butler that is exclusive for agricultural knowledge. Based on statistics on over 7 thousand questions, over 6 thousand answers, and over 5 thousand users, we first investigated the question repository and the user activities in Farm-Butler, and then we propose a graph model to analyze the user relationships. Finally, we applied our findings to build two machine-learned classifiers to predict active and popular experts.

Among our results, we find that Farm-Butler exhibits certain characteristics that are often observed in online social networks, for example, the highly-skewed content popularity. In addition, we find a number of fascinating distinctions in Farm-Butler. First, the users are most active during the summer time and the questions are mostly raised during the daytime. Secondly, on average, each farmer has raised 18.18 questions whereas each expert has replied 431 questions, indicating that the experts are more active than the farmer. Thirdly, in the graph model, the in-degree distribution of the farmers and the experts are similar, indicating that experts not only help others but also learn from the community. We leave a further analysis on the motivations of users taking different roles as our future work.

Acknowledgements

This work was partially supported by the National Science Foundation for Young Scholars of China (NSFYSC) No. 61502500 and Beijing Natural Science Foundation (BNSF) No. 4164090.

[References]

[1] Adamic, L. A., J. Zhang, E. Bakshy, and M. S. Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows something. *April 21-25, 2008:* In *Proc. 17th International Conference on World Wide Web*, 665–674. Beijing, China.

[2] Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1): 5–32.

[3] Ding, Y., Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, and A. Ghose. 2011. Broadcast yourself: understanding YouTube uploaders. *November 2-4, 2011: Proceeding of the 2011 ACM SIGCOMM on Internet Measurement Conference*, 361–370. Berlin, Germany.

[4] Figueiredo F., F. Benevenuto, and J. M. Almeida. 2011. The tube over time: characterizing popularity growth of youtube videos. February 9-12, 2011: *Proceeding of the 4th ACM International conf. on Web Search and Data Mining*, 745–754. Hong Kong, China.

[5] Harper, F. M., D. Moy, and J. A. Konstan. 2009. Facts or friends?: distinguishing informational and conversational questions in social Q&A sites. April, 4-9, 2009: *Proceeding of the SIGCHI Conference on Human Factors in Computing Systems*, 759–768. Boston, MA, USA.

[6] Hseih, G., and S. Counts. 2009. Mimir: a market-based real-time question and answer service. April 4-9, 2009: *Proceeding of the SIGCHI Conference on Human Factors in Computing Systems*, 769–778. Boston, MA, USA.

[7] Jia, A. L., S. Shen, D. H. Epema, and A. Iosup. 2016. When game becomes life: The creators and spectators of online game replays and live streaming. *ACM Transaction on Multimedia Computing, Communications, and Applications*, 12(4): 47.

[8] Morris, M. R., J. Teevan, and K. Panovich. 2010. What do people ask their social networks and why? A survey study of status message Q&A behavior. April 10-15, 2010, *Proceeding of the SIGCHI Conference on Human Factors in Computing Systems*, 1739–1748. Atlanta, GA, USA.

[9] Myers, S. A., A. Sharma, P. Gupta, and J. Lin. 2014. Information network or social network?: the structure of the twitter follow graph. April 7-11, 2014: *Proceeding of the 23rd International Conference on World Wide Web*, 493–498. Seoul, South Korea.

[10] Pal, A., S. Chang, and J. A. Konstan. 2012. Evolution of experts in question answering communities. *June 4-8, 2012: Proceeding of the 6th International AAAI Conference on Weblogs and Social Media*, 274–281. Dublin, Ireland.

[11] Panovich, K., R. Miller, and D. Kargerk. 2012. Tie strength in question & answer on social network sites. February 11-15, 2012: *Proceeding of the ACM 2012 Conference on Computer Supported Cooperative Work*, 1057–1066. Seattle, Washington, USA.

[12] Paul, S. A., L. Hong, and E. H. Chi. 2011. Is Twitter a good place for asking questions? a characterization study. July17-21, 2011: *Proceeding of the 5th International Conference on Weblogs and Social Media*, 578–581. Barcelona, Catalonia, Spain.

[13] Rodrigues, E. M., N. Milic-Frayling, and B. Fortuna. 2008. Social tagging behaviour in community-driven question answering. In *IEEE/WIC/ACM International Conf. on Web Intelligence and Intelligent Agent Technology*, vol. 3, 112–119. Sydney, Australia, 9-12 December.

[14] Shah, C., and J. Pomerantz. 2010. Evaluating and predicting answer quality in community QA. *July 19-23, 2010: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 411–418. Geneva, Switzerland.

[15] Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684): 440.