# User Churn in Focused Question Answering Sites: Characterizations and Prediction

Jagat Pudipeddi
Stony Brook University
jpudipeddi@cs.stonybrook.edu

Leman Akoglu
Stony Brook University
leman@cs.stonybrook.edu

Hanghang Tong
The City College of New York
tong@cs.ccny.cuny.edu

## ABSTRACT

Given a user on a Q&A site, how can we tell whether s/he is engaged with the site or is rather likely to leave? What are the most evidential factors that relate to users churning? Question and Answer (Q&A) sites form excellent repositories of collective knowledge. To make these sites self-sustainable and long-lasting, it is crucial to ensure that new users as well as the site veterans who provide most of the answers keep engaged with the site. As such, quantifying the engagement of users and preventing churn in Q&A sites are vital to improve the lifespan of these sites.

We study a large data collection from `stackoverflow.com` to identify significant factors that correlate with newcomer user churn in the early stage and those that relate to veterans leaving in the later stage. We consider the problem under two settings: given (i) the first $k$ posts, or (ii) first $T$ days of activity of a user, we aim to identify evidential features to automatically classify users so as to spot those who are about to leave. We find that in both cases, the time gap between subsequent posts is the most significant indicator of diminishing interest of users, besides other indicative factors like answering speed, reputation of those who answer their questions, and number of answers received by the user.

## Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; I.5.2 [**Design Methodology**]: Feature evaluation and selection

## Keywords

user churn; churn prediction; feature extraction; Q&A sites

## 1. INTRODUCTION

Online Question and Answer (Q&A) sites, such as StackOverflow, Yahoo! Answers, Quora, Baidu Knows (China), Naver (Korea), etc. are excellent platforms for satisfying information needs of Internet users in the form of well-crafted questions and well-rounded answers. These platforms often go beyond "asking Google" in asking questions, since the questions range from subjective ones to technical ones that are hard to obtain answers to by simple search engine querying. Moreover, the answers received are often beyond simple page results, rather, knowledgeable users tend to give detailed and direct answers to the questions. As such, Q&A sites are more of the online equivalent of a room full of experts where users can walk in and ask questions which other users might have answers to. Therefore, they are important crowdsourced knowledge repositories for millions of Internet users seeking information on the Web.

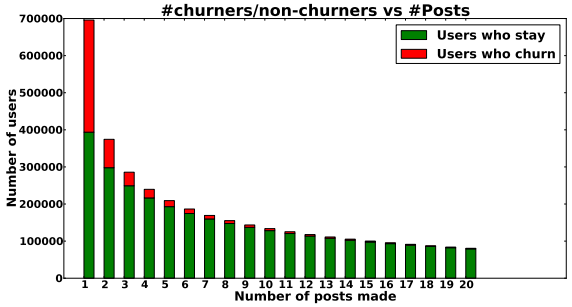


**Figure 1: Histogram of churning and staying users by post count (up to 20) in StackOverflow. User churn is an issue, with a large fraction of users churning after only a few posts.**

While being extremely important knowledge sources as alternatives to search engines such as Google and online encyclopedias such as Wikipedia, Q&A sites come with their own problems and challenges. Since a Q&A site's popularity is based on the breadth of the questions and answers provided, it is important for the site to make sure that users who post questions look at the site as reliable and can reach high-quality content that they are looking for as efficiently as possible while motivating those who provide answers to continue doing so. Thus, one of the main challenges has to do with user engagement: in order to make these sites self-sustainable, it is crucial for the site owners to keep the users engaged in asking and answering questions. In particular, the users should be well motivated to provide answers for reasons beyond monetary benefits, since almost none of the popular Q&A sites pay users to answer questions. One solution to this challenge involves a reputation system which offers the answerers various forms of virtual rewards.

However, user engagement often goes beyond virtual rewards and user churn is certainly a big problem that Q&A sites face. For example see Figure 1, which shows the fraction of StackOverflow users churning at various stages. We notice the skewed power-law-like distribution, showing


*WWW'14 Companion,* April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576965 .

that most users post only a few posts (in fact, 94% of the users have less than or equal to 20 posts), and more importantly notice that a large number of users (>350,000 in 4 years) leave the site after 1-2 posts.

One may think of numerous contributing factors to user churn. For example, possible (causal) factors related to question askers' leave may be untimely/no answers or low-quality answers received, while those contributing to answerers' leave may be high competition with other answerers or low-reward scores received for their answers. Moreover, there may be (not necessarily causal) signals in the system that point to a user's churn, such as decrease in frequency or amount of activity.

Our goal is to identify those factors that correlate with user churn in Q&A sites and automatically spot those users who are likely to leave. We formulate the problem as a classification task under two settings and answer the following questions: (i) given the first $k$ posts of a user, or (ii) given the first $T$ days activity of a user, how can we predict whether the user is about to churn? We use a large collection of data[1] publicly shared by StackOverflow to perform our analysis, while our framework is quite general for spotting user churn in other Q&A sites as well. Our main contributions are:

1. *Evidential features:* We explore a long list of potentially correlated features with user churn, which we organize under *nine* groups, in particular those related to time, frequency, quality, consistency, speed, gratitude, competitiveness, content, and finally knowledge level, and identify strongly indicative features. We also study how the discriminative power of a feature varies across various settings, i.e. for changing $k$ and $T$.
2. *User churn prediction:* Using the identified features, we learn classifiers to predict the likelihood of a given user to churn. Notably, we focus on two groups of users both of which are important for the Q&A sites; newbies with 1-5 posts who need to be nurtured into experts, and experts with more than 15 posts who provide most answers. Our analyses characterize user churn and achieve up to 74% prediction accuracy.

To the best of our knowledge, this is the first study of user churn in Q&A sites that analyzes various settings (varying number of posts and period of activity) and that takes both types of users (newbies and experts) into account.

## 2. Q&A SITES AND DATA DETAILS

The questions on Q&A sites can be of a wide variety such as finance, music, travel, how-to's, etc. Morris *et al.* [11] studied the kinds of topics for which people turn to online sites to seek answers to and found *technology* to be the top contender, for the wide array of problems people face when using technology such as programming tools or languages.

StackOverflow is a *technology-focused* question and answer site, where people ask specific software engineering and programming questions and others provide answers. Example questions include *"How to do in Java what in C++ is changing an overridden method visibility?"*, and *"How to avoid using for-loops with numpy?"*.

Our data[1] consists of site activity from July 31, 2008 to July 31, 2012, during which $\sim$ 3.4 million questions were posted, of which 91.3% were answered within a median time of 16 minutes. To achieve engagement, the site is designed such that users obtain reputation points, badges, and increasingly powerful tools as they post high quality content.[2]

While we focus on StackOverflow for analyzing user churn, other popular Q&A sites like Yahoo! Answers and Quora follow a similar model. Thus our work could be generalized to such sites. It remains as future work to cross-validate our findings empirically across sites.

## 3. RESEARCH QUESTIONS

We aim to identify the intrinsic factors and signals that cause a user to stop posting and use those signals to automatically determine likely-to-churn users in Q&A sites. Some of the extrinsic factors like the user losing interest in online activities manifest in the form of intrinsic signals on the site like increase in temporal inter-post gap, i.e. decrease in frequency of activity. Other factors that likely affect churn can be thought as related to speed and quality of answers, points rewarded, etc. Of course, not all factors can be directly observed or inferred from the available site information, like job loss or end of college studies, which may make the users to leave abruptly, and are factors that are hard to account for. Nevertheless, we aim to study a long list of potential factors to identify those that provide the strongest signals for churn.

The list of questions we aim to answer are as follows.

1. What are the intrinsic factors and signals that make a new user leave after a certain number of posts?
2. What makes a prolific user who has been posting significantly more posts than an average user leave after a certain number of posts?
3. Are the correlated factors common across these two groups of users (i.e., newcomer vs. prolific users)? If not, how do the evidential factors vary?
4. How well can we predict whether a user is likely to churn using the identified evidential features?

In order to answer these questions, we study the problem under two settings:

***Task 1.***

**Given** the first $k$ posts (questions and answers) of a user,

***Task 2.***

**Given** the first $T$ days of site activity of a user,

**Predict** how likely it is that the user will churn (i.e., will have no activity for the next 6 months).

We perform the above tasks for varying $k$ and $T$. In particular, we consider $1 \leq k \leq 5$ and $16 \leq k \leq 20$ as various number of posts. Although the vast majority of churning users leave within 5 posts (see Figure 1) and thus prediction of churn is potentially more beneficial in early stages, we also consider in our study the prediction of churn for users with more posts. The reason is that the first group of users include newbies with limited experience, whereas the second group includes more experienced, prolific users with more activity (although fewer in count) whose churn would also hurt the site. In other words while newcomers are large in terms of quantity, the veterans are potentially better in terms of experience and hence quality. Similarly for varying $T$, we consider $T = \{7, 15, 30\}$ days.

[1]StackOverflow data: `http://blog.stackoverflow.com/category/cc-wiki-dump/`, with more than one million users and spanning four years of activity. See §2 for data details.

[2]`http://meta.stackoverflow.com/questions/7237/how-does-reputation-work`

# 4. EVIDENTIAL FEATURES

## 4.1 Feature Description

As one of our main contributions, we construct and study a long list of potentially indicative features in churn, which we categorize into *nine* categories. We list all the features and their corresponding categories in Table 1. Note that the underlined features are used only in one task and not the other. For example, *num_posts* is only used in *Task 2* where we are given the first $T$ days activity of a user. This feature is uninformative for *Task 1*, where the number of posts of a user is fixed to $k$ for all users.

1. *Temporal* features are based on the time gaps between user activities. This captures the user's posting pattern, which potentially indicates the user's gradual change in interest.
2. *Frequency* features represent how often the user posts and whether s/he posts a question or an answer. This is based on the observation by [11] that churn probability decreases with increase in number of answers.
3. *Quality* features capture the quality of a user's posts, as reflected by the reputation/reward scores they receive for those posts.
4. *Consistency* features capture the consistency in quality of posts. Our insight is: the more consistent a user's posts are, the lower the chances of their churning.
5. *Speed* features represent how quick a user is in responding to another user's question. This is an indirect measure of the user's enthusiasm and intuitively negatively correlates with user churn.
6. *Gratitude* features represent how other users explicitly express gratitude on the user's posts. Intuitively, this is one of the measures of user's gratification and hence correlates negatively with churn.
7. *Competitiveness* features capture the user's will to provide higher quality answers when compared to other peers who answer the same question. We believe that the gratification obtained from an answer depends not only on the user's answer but also on the (number of) answers other users have provided.
8. *Content* features use the content of the posts, and are based on the observation [18] that at $k$=1, longer questions correlate with lower chance of churning.
9. *Knowledge Level* features capture how useful the user's knowledge is to the StackOverflow community. This is an indirect measure of community fit. For instance, if the user receives an answer for their question in a short time, it could mean that their domain interests match with frequent users of StackOverflow. Since higher community relevance leads to higher gratification, likely that it negatively correlates with churn.

## 4.2 Feature Analysis

Here we analyze our data for several features across users who churn and those who stay, and show their potential discriminative power in separating the two classes of users.

We start with our most significant feature: temporal gaps between user posts. Figure 2 shows the time gaps with various number of posts $k$, $2-5$ (left) and $17-20$ (right). We notice that the gaps keep increasing till a user churns, indicating the increasing infrequency in churning users' activity. On the other hand, the gaps are relatively stable for those

**Table 1: List of evidential features we identified relating to user churn, grouped into 9 categories. Underlined ones are used in only one of the tasks.**

| **Temporal** |
|---|
| ***gap1***: Time gap between account creation and first post |
| ***gapK***: *Task 1.* Time gap between $(k-1)^{th}$ post and $k^{th}$ post for each possible $k \leq K$ |
| ***last_gap***: *Task 2.* Time gap between the last post and the post before that |
| ***time_since_last_post***: *Task 2.* Time elapsed between the last post made and the observation deadline |
| ***mean_gap***: *Task 2.* Average time gap between posts made during the observation period |
| **Frequency** |
| ***num_answers***: Number of answers |
| ***num_questions***: Number of questions |
| ***ans_que_ratio***: Ratio of #answers to #questions |
| ***num_posts***: *Task 2.* Number of posts |
| **Quality** |
| ***ans_score***: Reputation score obtained per answer given |
| ***que_score***: Reputation score obtained per question asked |
| **Consistency** |
| ***ans_stddev***: Standard deviation of the reputation scores obtained for the answers |
| ***que_stddev***: Standard deviation of the reputation scores obtained for the questions |
| **Speed** |
| ***answering_speed***: Inverse of the time gap between a question being posted and the user answering it |
| **Gratitude** |
| ***ans_comments***: Average #comments made on the user's answer |
| ***que_comments***: Average #comments made on the user's question |
| **Competitiveness** |
| ***relative_rank_pos***: Average of total number of answers for a question divided by the rank of user's answer |
| **Content** |
| ***ans_length***: Average length of an answer |
| ***que_length***: Average length of a question |
| **Knowledge Level** |
| ***accepted_answerer_rep***: Mean reputation of the user whose answer was accepted |
| ***max_rep_answerer***: Mean reputation of the user who had the maximum reputation among all those who answered a question |
| ***num_que_answered***: Number of questions posted by the user that got answered |
| ***time_for_first_ans***: Time taken for the arrival of the first answer to a question. |
| ***rep_questioner***: Mean reputation of the user whose question was answered. |
| ***rep_answerers***: Mean reputation of the users who answered the question. |
| ***rep_co_answerers***: Mean reputation of the users who answered the same question as the control user |
| ***num_answers_recvd***: Mean number of answers received for every question the user posts |

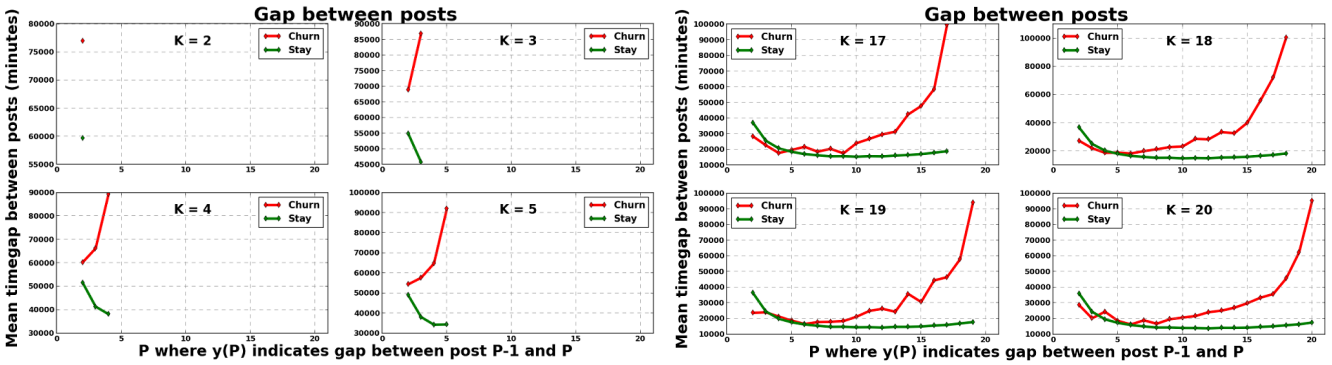


**Figure 2: For a user who churns, gap between consecutive posts keeps increasing. Gaps for those who stay are much lower, and stabilize around 20,000 minutes, indicating routine posting activity in every ≈2 weeks.**

who stay and in fact stabilizes at around 14 days showing routine posting activity of the staying users. Notably the difference between the gaps among the two classes of users increases, which implies that as $k$ increases the discriminating power of earlier posts reduces while the most recent ones become more and more relevant.

Next in Figure 3, we show the churn probability as a function of number of answers, for users with number of questions changing from 0 to 5. We clearly see that the more a user answers, and the more questions s/he asks, the less likely s/he is to churn. For users with the same number of answers, those with more questions tend to stay longer. This difference vanishes after about 5 answers, at which point the probability of churn drops with the same rate that is irrespective of the number of questions asked.

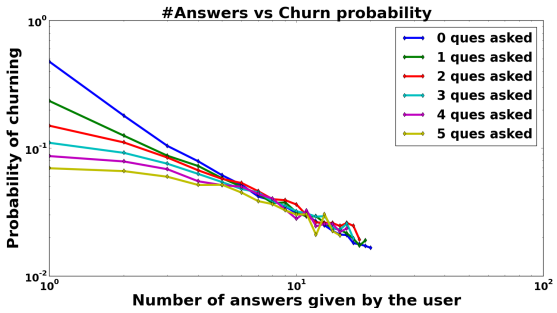


**Figure 3: The probability of churning for a user decreases the more answers s/he provides. It is even lower if s/he asks more questions alongside.**

In Figure 4, we observe that the longer a question poster has to wait for an answer, the higher the probability of churning. This can be explained by the negative correlation between waiting time (shown in minutes) and user satisfaction. We also notice that this probability is highest for $k$=1 and the difference reduces as $k$ increases.

Our data analysis also reveals that users who stay have a higher answer/question ratio, are speedier in answering questions, and provide higher quality answers than those users who churn (plots omitted due to page limit).

# 5. CHURN PREDICTION

Having identified evidential features that are indicative of user churn, we turn to exploiting them for churn prediction.

We consider users who do not post on the site for at least 6 months as having stopped using the site, and treat them

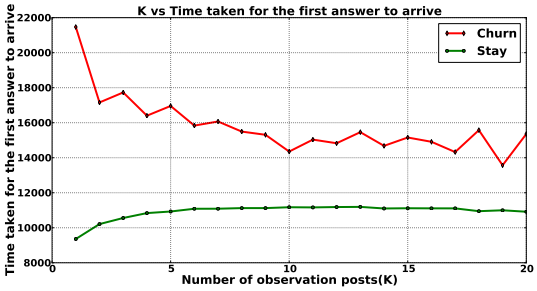


**Figure 4: The more the time taken for a user to receive an answer, the lesser the satisfaction level and the more the chances of churning.**

as the churners. Less than 13% of the users have had a gap of more than 6 months between their posts. We excluded from our study all the users who posted within 6 months prior to July 2012 (end date of data), since we cannot infer from the data if they posted in the future beyond that time.

For *Task 1*, the training set is obtained the following way: For each $k$, we select the users who did not post for at least 6 months from their $k^{th}$ post and users who created at least one post within the 6 months. For *Task 2*, we include in the training set those users who did not post for at least 6 months from $T$ days after account creation, and users who created at least one post within the 6 months, for each $T$.

As the number of users who churn reduces relative to the users who stay as $k$ or $T$ increases (see Figure 1), we adopt a similar strategy (under-sampling) in the imbalanced classification community [3] to make the two classes balanced, so that the numbers of users who stay and churn are equal in the training set. As such, the baseline accuracy is 50%.

For each task, we report 10-fold cross validation accuracies for various $k$ and $T$ as our performance measure.

## 5.1 Prediction Results

For churn prediction, there are several possible models that one could employ. In Tables 2 and 3, we show the prediction accuracy of four different classification models on our two tasks for various $k$ and $T$, respectively. Among the four, decision tree (DT) and SVM with the RBF kernel have nonlinear decision boundaries, while (regularized) logistic regression (LR) and SVM with the linear kernel aim to split the data with a linear hyperplane.

For the experiments where we vary $k$, decision trees perform the best. Especially when observation data is scarce

**Table 2: Performance on *Task 1* of various classifiers with changing $k$. Decision trees perform the best, especially for limited data, i.e., when $k$ is small (the difference diminishes with $k$).**

| $k$ (posts) | Decision Tree | SVM (Linear) | SVM (RBF) | Logistic Regression |
|---|---|---|---|---|
| 1 | **72.6** | 60.9 | 61.2 | 61.1 |
| 2 | **67.1** | 58.6 | 59.4 | 58.7 |
| 3 | **64.4** | 59.5 | 60.2 | 59.5 |
| 4 | **65.0** | 60.6 | 61.2 | 60.7 |
| 5 | **65.2** | 62.4 | 63.1 | 62.7 |
| 16 | **69.4** | 68.5 | 69.0 | 69.3 |
| 17 | **69.7** | 68.9 | 68.9 | 69.4 |
| 18 | 70.3 | 69.7 | **70.4** | 70.3 |
| 19 | 69.3 | 69.2 | 69.2 | **69.6** |
| 20 | **71.2** | 69.7 | 69.9 | 70.1 |

**Table 3: Performance on *Task 2* of various classifiers with changing $T$. Decision trees perform the best (performance gets better as $T$ increases, i.e. when more data is available).**

| $T$ (days) | Decision Tree | SVM (Linear) | SVM (RBF) | Logistic Regression |
|---|---|---|---|---|
| 7 | **70.6** | 67.0 | 67.4 | 67.0 |
| 15 | **72.2** | 69.9 | 70.3 | 70.1 |
| 30 | **74.1** | 72.5 | 73.3 | 72.7 |

(i.e., for small $k = 1 \dots 5$), other classifiers perform much worse than DT. For $k = 16 \dots 20$, their performances come closer to that of DT. One hypothesis for this behavior is that as $k$ increases, the decision boundaries smoothen out enough to allow SVMs and LR to perform as well as DT does. For the experiments where we vary $T$, DT again achieves the best accuracies, and performance gets better as $T$ increases.

Since they performed the best among several prediction models, we proceed with the DT classifiers for fine grained feature characterization in the remainder of this section.

## 5.2 Feature Analysis

Next we aim to quantify the importance of each feature category in isolation. To do so, we performed classifications using only the features of each category.

We give their prediction results in Figure 5, for *Task 1* (left) and for *Task 2* (right). We obtain the best accuracy of 72.6% at $k$=1 for *Task 1*, and 74.1% at $T$=30 for *Task 2* using all the features. In fact, for every $k$ and every $T$ in *Task 1* and *Task 2* respectively, the model learned on all features gives the best result. Among the models trained on individual feature categories, the one that uses only the *temporal features* provides the best accuracy which gets quite close to that of using all features (69.7% at $k$=1 and 73.1% at $T$=30). Models learned using three other feature categories, namely *knowledge level*, *content*, and *frequency* rank the next best. We also realize that the predictive power of each feature category for newbies and veterans (i.e., small vs. large $k$ and $T$) is quite comparable, suggesting the generalized potential of feature groups across these user types.

To further highlight the predictive power of temporal gap features in user churn, we demonstrate in Table 4 the prediction accuracies when (i) all features are used, (ii) only all temporal gap features are used, and finally (iii) only the single last-gap feature is used. We observe that the models learned with only the temporal gap features achieve accuracies quite close to (if not better than) what we obtain with all features. Using the last gap feature by itself provides reasonably high performance.

**Table 4: Temporal gap features provide the highest boost in accuracy. Using only the temporal features achieves similar accuracy to that of all features.**

| $k$ | All Features | Only ***gapK*** (Temporal Gaps) | Only ***last_gap*** (Last-Gap) |
|---|---|---|---|
| 1 | **0.726** | 0.697 | 0.697 |
| 3 | **0.644** | 0.611 | 0.566 |
| 5 | **0.652** | 0.635 | 0.608 |
| 8 | **0.676** | 0.662 | 0.636 |
| 10 | **0.675** | 0.670 | 0.649 |
| 13 | 0.680 | **0.682** | 0.655 |
| 15 | 0.691 | **0.694** | 0.666 |
| 18 | 0.703 | **0.706** | 0.679 |
| 20 | 0.712 | **0.713** | 0.688 |

We also notice that the temporal gap models slightly outperform the ones with all the features for larger $k$. This may be explained by model complexity: with $k$ increasing, the number of features increases for both models. As the model (search) space becomes larger for the (latter) complete model than for the (former) temporal model, the greedy decision tree algorithm [14] is more likely to land on local optimum during the search for the complete model.

These results corroborate our observations in Figure 2 which demonstrated that the time gaps of churning users kept increasing over time, while stabilizing for non-churning ones, and thus creating separation between the two groups.

In summary, for both of our tasks, the temporal features perform notably well. The cost of computing these features is low, thus one can argue that these are suitable and practical for user churn prediction in the real-world. Other feature categories can be used to boost the accuracy further as required, at a cost of computational power.

# 6. RELATED WORK

There exist several works studying the user lifespan in online Q&A sites. Yang *et al.* analyzed three large sites from three countries to understand the predictive patterns in participation lifespans of users [18]. Later, they studied the cultural effects in people's behavior [17], such as the motivating factors for asking and answering questions, which showed differences across countries. Several other studies focused on newcomers' retention [9, 8, 4], showing that the first interaction is critical for sustaining a large number of future users. Arguello *et al.* studied user communities to understand the contributing factors to success in their ability to respond to and retain active participants [2]. Those works study users' overall behavior, and come up with general statements like "longer answers received make askers to stay". In our study we explore a wide variety of potential factors and learn prediction models, that can do individualized predictions for both the newcomers and the veterans.

Researchers have also studied sustainability in other settings; such as social networks [5, 13, 12] and telecommunication networks [15, 7] where the goal is to analyze the cascades of users leaving the network. Different from users leaving Q&A sites, the type of churn in such networks has more of a social context, where one's friends leaving impacts

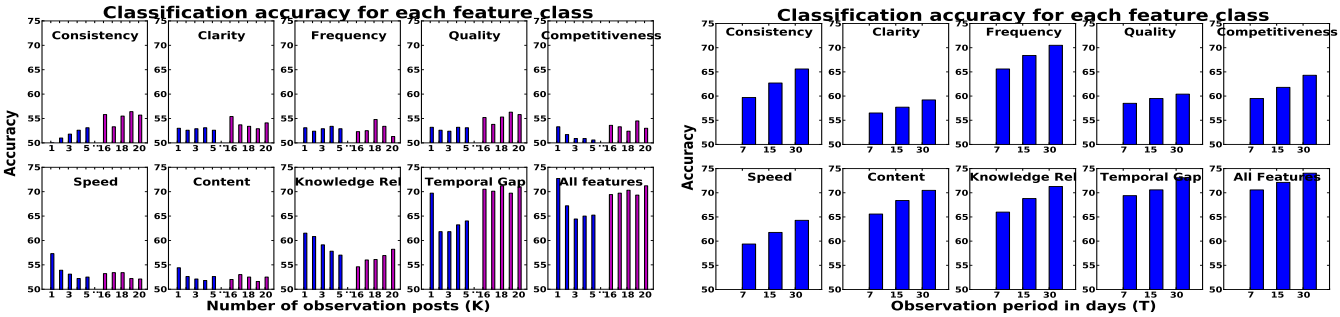


**Figure 5: Churn prediction accuracy when features from each category are used in isolation, (left) as $k$ varies and (right) as $T$ varies. Temporal gap features alone provide competitive accuracy that is close to that of all features. Accuracies tend to increase with more observational data, that is for larger $k, T$. For both tasks, the predictive power of each feature category for newbies and veterans (i.e., small vs. large $k, T$) is comparable, where temporal gap features are the most significant for both user types.**

his/her leave. In Q&A sites, however, the correlations are more intricate than direct relations to other users in the site.

Wang *et al.* studied Quora to understand the impact of its site design and organization, specifically the underlying connectivity structures among its members, related questions, and user-question topic relations, on the growth and quality of its knowledge base [16]. Different from earlier behavioral studies, this work focuses on the design aspects of the Q&A sites and their effects on user engagement.

Other related works include the analysis of the quality and value of questions and answers in Q&A sites. Anderson *et al.* analyzed the factors that contribute to the long-term value of questions in StackOverflow [1]. Harper *et al.* studied the predictors of answer quality with respect to two dimensions; the site characteristics such as the type and organization of communities and experts, and the question characteristics including strategies such as thanking in advance and showing prior effort [6]. Liu *et al.* looked into inferring the satisfaction of question askers [10]. These indicators are potentially useful predictors of churn as well, which could be incorporated in future studies of churn.

# 7. CONCLUSION

In this work we studied a large collection of StackOverflow data with the goal of identifying factors that correlate with user churn in Q&A sites. Data analysis showed that several factors, such as time gaps between user posts, provide strong evidence for churn. We built a long list of such evidential features, organized into nine categories. We used these informative features to learn churn prediction models. Experiments demonstrated that exploiting site-level information could help spot likely user churn in Q&A sites.

## Acknowledgements

We thank the reviewers for helping us improve our manuscript. This material is based on work supported by the Army Research Office under Contract No. W911NF-14-1-0029 and Stony Brook University Office of Vice President for Research. Any findings and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

# 8. REFERENCES

[1] A. Anderson, D. P. Huttenlocher, J. M. Kleinberg, and J. Leskovec. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *KDD*, pages 850–858, 2012.

[2] J. Arguello, B. S. Butler, E. Joyce, R. Kraut, K. S. Ling, , and X. Wang. Talk to me: foundations for successful individual-group interactions in online communities. In *CHI*, 2006.

[3] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *DAMI*, pages 875–886. Springer, 2010.

[4] G. Dror, D. Pelleg, O. Rokhlenko, and I. Szpektor. Churn prediction in new users of Yahoo! answers. In *WWW*, pages 829–834, 2012.

[5] D. Garcia, P. Mavrodiev, and F. Schweitzer. Social resilience in online communities: The autopsy of friendster. *CoRR*, abs/1302.6109, 2013.

[6] F. M. Harper, D. R. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online Q&A sites. In *CHI*, pages 865–874. ACM, 2008.

[7] Y. Huang, B. Q. Huang, and M. T. Kechadi. A rule-based method for customer churn prediction in telecommunication services. In *PAKDD*, 2011.

[8] E. Joyce and R. E. Kraut. Predicting continued participation in newsgroups. *J. Computer-Mediated Communication*, 11(3):723–747, 2006.

[9] C. Lampe and E. W. Johnston. Follow the (slash) dot: effects of feedback on new members in an online community. In *GROUP*, pages 11–20, 2005.

[10] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *SIGIR*, pages 483–490, 2008.

[11] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: a survey study of status message Q&A behavior. In *SIGCHI*, 2010.

[12] B. Ngonmang, E. Viennet, and M. Tchuente. Churn prediction in a real online social network using local community analysis. In *ASONAM*, 2012.

[13] R. J. Oentaryo, E.-P. Lim, D. Lo, F. Zhu, and P. K. Prasetyo. Collective churn prediction in social network. In *ASONAM*, pages 210–214, 2012.

[14] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[15] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *SDM*, pages 732–741, 2010.

[16] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the social crowd: an analysis of Quora. In *WWW*, 2013.

[17] J. Yang, M. R. Morris, J. Teevan, L. A. Adamic, and M. S. Ackerman. Culture matters: A survey study of social Q&A behavior. In *ICWSM*, 2011.

[18] J. Yang, X. Wei, M. S. Ackerman, and L. A. Adamic. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. In *ICWSM*, 2010.