The Social World of Content Abusers in Community Question Answering

Imrul Kayes
Computer Science and Engineering
University of South Florida
Tampa FL, USA
imrul@mail.usf.edu

Nicolas Kourtellis
Yahoo Labs
Barcelona, Spain
kourtell@yahoo-inc.com

Daniele Quercia
Yahoo Labs
Barcelona, Spain
dquercia@yahoo-inc.com

Adriana Iamnitchi
Computer Science and Engineering
University of South Florida
Tampa FL, USA
anda@cse.usf.edu

Francesco Bonchi
Yahoo Labs
Barcelona, Spain
bonchi@yahoo-inc.com

ABSTRACT

Community-based question answering platforms can be rich sources of information on a variety of specialized topics, from finance to cooking. The usefulness of such platforms depends heavily on user contributions (questions and answers), but also on respecting the community rules. As a crowd-sourced service, such platforms rely on their users for monitoring and flagging content that violates community rules.

Common wisdom is to eliminate the users who receive many flags. Our analysis of a year of traces from a mature Q&A site shows that the number of flags does not tell the full story: on one hand, users with many flags may still contribute positively to the community. On the other hand, users who never get flagged are found to violate community rules and get their accounts suspended. This analysis, however, also shows that abusive users are betrayed by their network properties: we find strong evidence of homophilous behavior and use this finding to detect abusive users who go under the community radar. Based on our empirical observations, we build a classifier that is able to detect abusive users with an accuracy as high as 83%.

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Social Issues—Abuse and crime involving computers; J.4 [**Social and Behavioural Sciences**]: Sociology

Keywords

Community question answering; content abusers; flagging; crowdsourcing;

1. INTRODUCTION

Community-based Question-Answering (CQA) sites, such as Yahoo Answers, Quora and Stack Overflow, are now rich and mature repositories of user-contributed questions and answers. For example, Yahoo Answers, launched in December 2005, has more than one billion posted answers[1]. Quora, one of the fastest growing CQA sites has seen three times growth in 2013[2].

Like many other Internet communities, CQA platforms define community rules and expect users to obey them. To enforce these rules, published as community guidelines and terms of services, these platforms provide users with tools to flag inappropriate content. In addition to community monitoring, some platforms employ human monitors to evaluate abuses and determine the appropriate responses, from removing content to suspending user accounts.

To the best of our knowledge, this paper is the first to investigate the reporting of rule violations in Yahoo Answers (*YA*), one of the oldest, largest, and most popular CQA platforms. The outcomes of this study could aid human monitors with automate tools in order to maintain the health of the community. Our dataset contains about 10 million editorially curated abuse reports posted between 2012 and 2013. Out of the 1.5 million users who submitted content during the one-year observation period, about 9% of the users got their accounts suspended. We use suspended accounts as a ground truth of bad behavior in *YA*, and we refer to these users as *content abusers*.

We discover that, although used correctly, flags do not tell accurately which users should be suspended: while 32% of the users active in our observation period have at least one flag, only 16% of them are suspended during this time. Even considering the top 1% users with the largest number of flags, only about 50% of them deserve account suspension. Moreover, we see that users with lots of flags contribute positively to the community in terms of providing (even best) answers. Complicating an already complex problem, we find that 40% of the suspended users have not received any flags.

[1] http://www.yanswersbloguk.com/b4/2010/05/04/1-billion-answers-served/
[2] http://www.goo.gl/MfK83y

To reduce this large gray area of questionable behavior, we employ social network analysis tools in an attempt to understand the position of content abusers in the *YA* community. We learned that the follower-followee social network tunnels user attention not only in terms of generating answers to posted questions, but also in monitoring user behavior. More importantly, it turns out that this social network divulges information about the users who go under the community radar and never get flagged even if they seriously violate community rules. This network-based information, combined with user activity, leads to accurate detection of the "bad guys": our classifier is able to distinguish between suspended and fair users with an accuracy as high as 83%.

The paper is structured as follows. Section 2 discusses previous analysis of CQA platforms and the existing body of work on unethical behavior in online communities in general. Section 3 presents the *YA* functionalities relevant to this study and the dataset used. We introduce a deviance score in Section 4 that identifies the pool of bad users more accurately than the number of flags alone. Section 5 demonstrates that deviant users are not all bad: despite their high deviance score, in aggregate their presence in the community is beneficial. Section 6 shows the effects of the social network on user contribution and behavior. Section 7 shows the classification of suspended and fair users. We discuss the impact of these results in Section 8.

2. RELATED WORK

We collate past research on Community-based Question Answering (CQA) in four categories depending on whether it has dealt with content, users, new applications, or CQA communication networks.

Content. Research in this area has investigated textual aspects of questions and answers. In so doing, it has proposed algorithmic solutions to automatically determine: the quality of questions [16, 30] and answers [27, 1, 15], the extent to which certain questions are easy to answer [9, 26], and the type of a given question (e.g., factual or conversational) [13, 14].

Users. Research on CQA users has been mostly about understanding why users contribute content: that is, why users ask questions (askers are failed searchers, in that, they use CQA sites when web search fails [17]); and why they answer questions (e.g., they refrain from answering sensitive questions to avoid being reported for abuse and potentially lose access to the community [7]).

New applications. As for applications, research has proposed effective ways of recommending questions to the most appropriate answerers [25, 31], of automatically answering questions based on past answers [28], and of retrieving factual answers [4] or factual bits within an answer [33].

Communication networks. The communication networks behind CQA sites have been recently studied. More specifically, researchers have explored the relationship between content quality and network properties such as number of followers [32] and tie strength [23].

Bad behavior in online settings. Qualitative and quantitative studies of bad behavior in online settings have been done before including newsgroups [24], online chat communities [29], and online multiplayer video games [5].

Impact of bad behavior. A body of work also investigates the impact of the bad behavior. Researchers find that bad behavior has negative effects on the community and its members: it decreases community's cohesion [34], performance [10] and participation [6]. In the worst case, users who are the targets of bad behavior may leave or avoid online social spaces [6].

Research on CQA communication networks is quite recent, so it comes as no surprise that there has not been any work on how such networks mediate different types of behavior on CQA sites. This paper, for the first time, sheds light on bad behavior in CQA communities by studying Yahoo Answers, one of the largest and oldest such communities. It quantifies how Yahoo Answers' networks channel user attention, and how that results in different behavioral patterns that can be used to limit bad behavior.

3. YAHOO ANSWERS

After 9 years of activity, Yahoo Answers has 56M monthly visitors (U.S. only)[3]. The functionalities of the *YA* platform and the dataset used in this analysis are presented next.

3.1 The Platform

YA is a CQA platform in which community members ask and answer questions on various topics. Users ask questions and assign them to categories selected from a predefined taxonomy, e.g., *Business & Finance*, *Health*, and *Politics & Government*. Users can find questions by searching or browsing through this hierarchy of categories. A question has a title (typically, a short summary of the question), and a body with additional details.

A user can answer any question but can post only one answer per question. Questions remain open for four days for others to answer. However, the asker can select a best answer before the end of this 4-day period, which automatically *resolves* the question and archives it as a *reference* question. The best answer can also be rated between one to five, known as *answer rating*. If the asker does not choose a best answer, the community selects one through voting. The asker can extend the answering duration for an extra four days. The questions left unanswered after the allowed duration are deleted from the site. In addition to questions and answers, users can contribute comments to questions already answered and archived.

YA has a system of points and levels to encourage and reward participation[4]. A user is penalized five points for posting a question, but if she chooses a best answer for her question, three points are given back. A user who posts an answer receives two points; a best answer is worth 10 points.

A leaderboard, updated daily, ranks users based on the total number of points they collected. Users are split into seven levels based on their acquired points (e.g., 1-249 points: level 1, 250-999 points: level 2, ..., 25000+ points: level 7). These levels are used to limit user actions, such as posting questions, answers, comments, follows, and votes: e.g., first level users can ask 5 questions and provide 20 answers in a day.

[3] http://www.listofsearchengines.org/qa-search-engines

[4] https://answers.yahoo.com/info/scoring_system

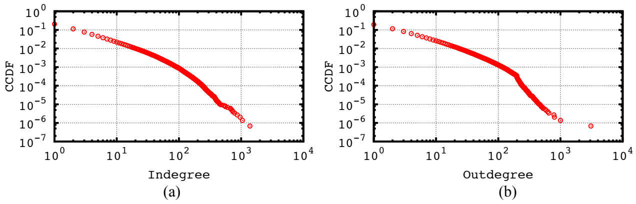
YA requires its users to follow the Community Guidelines that forbids users to post spam, insults, or rants, and the Yahoo Terms of Service [2] that limits harm to minors, harassment, privacy invasion, impersonation and misrepresentation, and fraud and phishing. Users can flag content (questions, answers or comments) that violates the Community Guidelines and Terms of Service using the "Report Abuse" functionality. Users click on a flag sign embedded with the content and choose a reason between violation of the community guidelines and violation of the terms of service. Reported content is then verified by human inspectors before it is deleted from the platform.

Users in *YA* can choose to follow other users, thus creating a follower-followee relationship used for information dissemination. The followee's actions (e.g., questions, answers, ratings, votes, best answer, awards) are automatically posted on the follower's newsfeed. In addition, users can follow questions, in which case all responses are sent to the followers of that question.

3.2 Dataset

We studied a sample of 10 million abuse reports posted between 2012 and 2013 originating from 1.5 million active users. These users are connected via 2.6 million follower-followee relationships in a social network (referred to as FF in this study) that has $165,441$ weakly connected components. The largest weakly connected component has 1.1M nodes (74.32% of the nodes) and 2.4M edges (91.37% of the edges). Out of the 1.5 million users, about 9% of the users have been suspended from the community.

Figure 1(a) and Figure 1(b) plot the complementary cumulative distribution function (CCDF) for the degree of followers (indegree) and followees (outdegree), respectively. The indegree and outdegree follow power-law distributions [3].



(a)

(b)

Figure 1: (a) Indegree distribution; (b) Outdegree distribution.

Along with the follower-followee social network, we built an activity network (AN) that connects users if they interacted with each other's content. In the AN network, nodes are users who answered other users' questions, directed edges point from the answerer to the asker, and edge weights show the number of answers provided over the source user to the questions posted by the destination user. The activity network has 1.2M nodes and 45M edges, thus being 141 times denser than the FF network.

4. FLAGGING IN YAHOO ANSWERS

In this section, we study whether flags (we use flags and abuse reports interchangeably) can be used as an appropriate proxy for content abuse. First, we investigate whether the flags reported from users are typically valid, i.e. if human inspectors remove the flagged content and further, how quickly this is done. Then, we explore how the flags can be used to detect content abusers.

4.1 Abuse Reports

YA is a self-moderating community; the health of the platform depends on community contributions in terms of reporting abuses. Besides participating by providing questions and answers, *YA* users also contribute to the platform by reporting abusive content. Reporters serve as an intermediate layer in the *YA* moderation process since these abuse reports are verified by human inspectors. If the report is valid, the content is promptly deleted.

To check if valid abuse reports are indeed an accurate sensor for the correct monitoring of the platform, we look at how soon a report is curated. Figure 2 shows the distributions of the time interval between the time when a content (question or answer) is posted and when it is deleted due to abuse reports. About 97% of questions and answers marked as abusive are deleted within the same day they are posted. All reported abusive questions and answers are deleted within three days of posting.

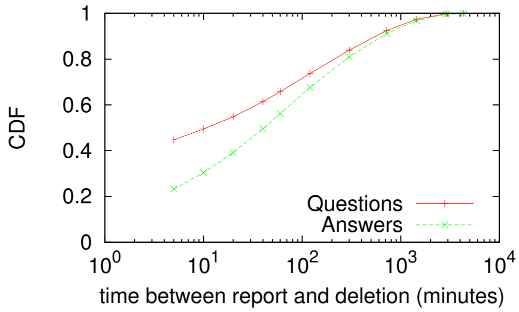


Figure 2: The CDF of the time delay between the posting of the content (questions or answers) and its deletion due to valid abuse reporting.

This result highlights two facts. First, that the users monitoring the platform act very quickly on content: within 10 minutes from being posted, 50% of the bad posts are reported. Second, that validation of abuse reports happens within 3 days (and in vast majority within a day). Hence, in our dataset, if there are abuse reports that did not have the chance of being curated yet and thus we do not consider them, those are too few to impact our analysis.

However, the abuse reporting functionality might be abused as well, due to several reasons. First, reporting is an easy and fast process, requiring only a few steps. Second, a user is not penalized for misreporting content abuse, perhaps in an attempt to not discourage users from exercising good citizenship. And third, independent of their level in the *YA* platform (that limits the number of questions and answers posted per day), users can report an unlimited number of abuses.

To check whether users abuse the abuse reporting functionality, we compare the number of flags received/reported with the number of validated flags received/reported per user. Figure 3 shows a correlation heat map of the flags received, flags received valid, flags reported and flags reported valid on questions for all contributors (results on answers are similar and are excluded for brevity). For questions (answers), we have a very high correlation between flags received by users and flags that are valid ($r = 0.90$ (0.87), $p < 0.01$) and between flags reported by users and that are valid ($r = 0.80$ (0.92), $p < 0.01$).

These high correlations indicate that, in general, users are not exploiting the abuse reporting functionality. When a user reports an abuse, it is very likely that the content is violating community rules. Another interesting finding from the correlation heat maps is that for both questions and answers, users have almost negligible or very weak correlation between the number of flags they reported that are valid and the number of flags they received that are valid. This hints that the good guys of the community are not bad guys at the same time: the users who correctly report a lot of content abuses are not posting abusive content themselves.

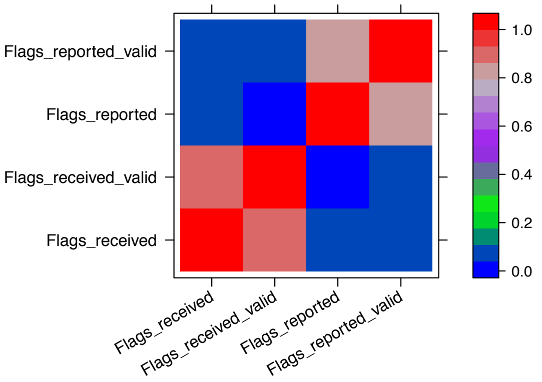


Figure 3: The Pearson correlation coefficient heat map of flags received, flags received valid, flags reported and valid flags reported on questions. All values are statistically significant (p-values <0.01).

4.2 Deviant Users

Given that flags are good proxies for identifying bad content, how should they be used to detect content abusers and thus determine which accounts to be suspended? Common wisdom might suggest that content abusers are those who receive a large number of flags. Of the top 1% flagged askers and answerers, we find 51.63% and 53.89%, respectively, are suspended. But finding a threshold on the number of flags received by a user is not likely to work accurately for content abuser detection: users with low activity who received flags for all their posts might go below this threshold. At the same time, highly active users may collect many flags even if for a small percentage of their posts, yet contribute significantly to the community.

This intuition motivated us to measure the correlation between a user's number of posts and the number of flags received. Indeed, we find that the correlation between the number of questions a user asks and the number of valid flags she receives from others is high ($r = 0.49$, $p < 0.05$). Similarly, the number of answers posted and the number of valid flags received per user are highly correlated ($r = 0.37$, $p < 0.05$). The distributions of the fraction of flagged questions and answers is shown in Figure 4. While about 27% users have more than 25% flagged questions, about 34% users have more than 25% flagged answers. Also, about 16% and 19% of users have more than 50% flagged questions and answers respectively.

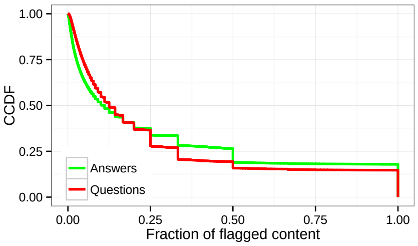


Figure 4: Distributions of fraction of flagged questions and answers.

So, instead of directly considering flags, we define a *deviance score* metric that indicates how much a user deviates from the norm in terms of received flags considering the amount of activity. Deviant behavior is defined by actions or behaviors that are contrary to the dominant norms of the society [8]. Although social norms differ from culture to culture, within a context, they remain the same and they are the rules by which the members of the community are conventionally guided.

We define the deviance score for a user u as the number of correct abuse reports (flags) she receives over the total content (question/answer) she posted, after eliminating the expected average number of correct abuse reports given the amount of content posted:

$$\text{Deviance}_{\text{Q/A}}(u) = Y_{Q/A,u} - \hat{Y}_{Q/A,u} \tag{1}$$

where $Y_{Q/A,u}$ is the number of correct abuse reports received by u for her questions/answers, and $\hat{Y}_{Q/A,u}$ is the expected number of correct abuse reports to be received by u for those questions/answers.

To capture the expected number of the correct abuse reports a user receives for questions/answers, we considered a number of linear and polynomial regression models between the response variable (number of correct abuse reports) and the predictor variable (number of questions/answers). Among them, the following linear model was the best in explaining the variability of the response variable.

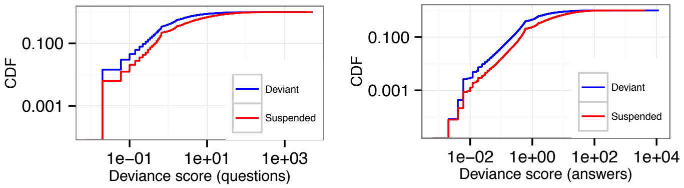
$$Y = \alpha + \beta X + \epsilon \tag{2}$$

where Y is the number of correct abuse reports (flags) received for the content, X is the number of content posts and ϵ is the error term.

In eq. (1), a positive deviance score reflects deviant users, i.e., those whose deviance cannot be only explained by their activity levels.

4.3 Deviance Score vs. Suspension

We have found 105,340 users with positive *question* deviance scores and 121,705 users with positive *answer* deviance scores. Among the users with positive question deviance score, 31,891 users (30.27%) have been suspended. Similarly, among the users with a positive answer deviance score, 37,633 users (30.92%) have been suspended. The CDF of suspended and deviant (but not suspended) users' deviance scores for both questions and answers is shown in Figure 5. In both cases, suspended and deviant users are visibly characterized by different distributions: suspended users tend to have higher deviance scores than deviant (not suspended) users. While this difference is visually apparent, we also ensure it is statistically significant using two methods: 1) the two-sample Kolmogorov-Smirnov (KS) test, and 2) a permutation test, to verify that the two samples are drawn from different probability distributions.



(a) (b)

Figure 5: The CDF of suspended and deviant users' deviance scores for (a) questions; (b) answers. Distributions are different with $p<0.001$ for both KS and permutation tests. For questions: $D = 0.22$, $Z = 46.04$. For answers: $D = 0.28$, $Z = 50.53$.

We also find that 63.94% of top 1% deviant question askers' and 64.77% of top 1% deviant answerers' accounts have been suspended. This hints that the higher deviance score a user has, the more likely (s)he is to be removed from the community. Figure 6 shows the probability of a user being suspended as a function of its rank in the community as expressed by deviance score and number of flags. We observe that the more deviant a user is, the more probable is that she will be suspended. Also, in all cases, deviance score shows a higher probability of suspension compared to the number of flags.

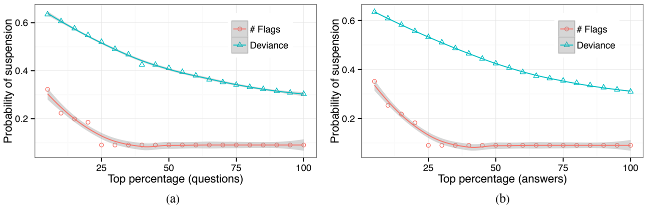


Figure 6: Probability of being suspended, given a user is within top x% of (a) question or (b) answer deviance scores and flags. 95% confidence interval area is shown.

These results show that the deviance score is a better metric for identifying the content abusers than the number of flags is by itself. However, both metrics fail to identify content abusers who go under the community radar. We found that about 40% of the suspended users had never been flagged for the abusive content they certainly posted, thus maintaining a negative deviance score. Thus, our investigation into user behavior in the *YA* community continues.

5. DEVIANT VS. SUSPENDED USERS

Despite the fact that deviance score better identifies the pool of suspended users, it is clearly an imperfect metric. On one hand, there are high deviance score users who are not suspended, despite the fact that the platform seems to be fairly quick in responding to abuse reports. On the other hand, there are "ordinary" users, according to the deviance score (i.e., with a negative deviance score) who are never reported for abusive content, yet get suspended. To better understand these two groups of users—deviant but not suspended and suspended but not flagged—we analyze in more detail their activity. Note that the two groups are disjoint (i.e., deviant users have received at least one flag).

Table 1: Descriptive statistics of the number of answers received by typical, deviant but not suspended, and suspended users per question.

Type	Min.	1st Qu.	Med.	Mean	3rd Qu.	Max.
Typical	1.00	1.00	2.00	4.36	5.00	1296.00
Deviant	1.00	5.00	11.00	17.96	22.00	1205.00
Suspended	1.00	1.00	4.00	8.67	9.00	1144.00

5.1 Deviance is Engaging

One of the success metrics of CQA platforms is *user engagement* [18], which can be measured by the number of contributions and by the number of users who respond to a particular content. Thus, we use the number of answers deviant users receive to their questions and the number of distinct users who respond to the deviant users' questions as measures of deviants' contribution to user engagement with the platform. To this end, for each category of users (typical, deviant but not suspended, and suspended) we randomly selected $500,000$ questions they asked. For each question, we extracted all answers received and also the users who answered those questions. Table 1 presents the statistics of the number of answers received per category of users.

Deviant users' questions get significantly more answers than typical users's questions get ($p_{ks} < 0.01$, $p_{perm} < 0.01$): on average, a question posted by a deviant user gets about 5 times more answers than the average question posted by a typical user. This difference is also seen in the CCDF of the number of answers received by typical, deviant and suspended users in Figure 7(a).

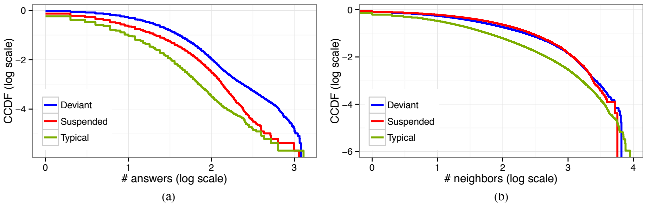


Figure 7: (a) CCDF of the number of answers received by the typical, deviant but not suspended, and suspended users on questions; (b) CCDF of the number of neighbors (distinct answerers) that typical, deviant but not suspended, and suspended users have.

Deviant users not only attract more answers, but also interact with more users than typical users do ($p_{ks} < 0.01$, $p_{perm} < 0.01$), as shown by Figure 7(b). This result from analyzing a random sample of 500,000 questions is confirmed when looking at the indegree of nodes in the activity network, which represents the number of users who answered that node's questions, as shown in Table 2 for typical and deviant users. Deviant askers have a higher number of neighbors than typical askers. An explanation might be, as shown in [13], that users who ask conversational questions tend to

Table 2: Descriptive statistics of the number of neighbors askers have in the Activity Network.

Type	Min.	1st Qu.	Med.	Mean	3rd Qu.	Max.
Typical	0.00	1.00	5.00	28.16	19.00	13270.00
Deviant	0.00	3.00	20.00	103.40	90.00	5698.00
Suspended	0.00	2.00	13.00	88.62	60.00	6576.00

have more neighbors (with whom the asker has interaction) than users who ask informational questions. This suggests that deviant users tend to ask more conversational questions, which engage a larger number of responders.

5.2 Deviance is Noisy

We observed that deviant users impact the *quantity* of content in the system. Do they impact *quality*, too? To address this question, we look at the following ratio of the best answers in the total number of answers submitted per user.

$$\text{Ratio of best answers}_u = \frac{\#\ \text{best answers from u}}{\#\ \text{total answers from u}} * 100 \tag{3}$$

Figure 8 shows the CDF of the ratio of best answers for different classes of users: 1) typical, 2) deviant but not suspended, and 3) suspended. The results show that users who are moderately deviant but did not get suspended have higher ratio of best answers than suspended users ($p_{ks} < 0.01$, $p_{perm} < 0.01$), but lower than that of typical users ($p_{ks} < 0.01$, $p_{perm} < 0.01$).

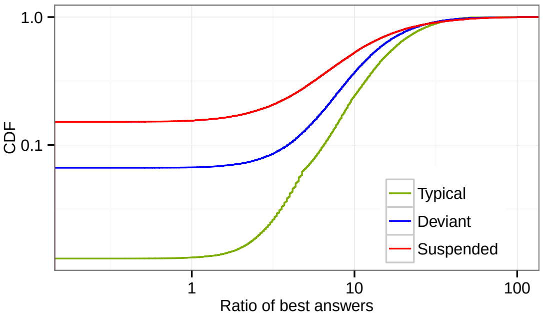


Figure 8: CDF of the ratio of best answers for typical, deviant but not suspended and suspended users.

To conclude, it turns out that while deviant users are beneficial in terms of platform success metrics, as they increase user engagement by attracting more answers and attracting more users who answer their questions, they do not contribute more than the norm-following users in terms of content quality.

5.3 The Suspended but Not Flagged Users

While the results above show how the deviant users differ from the suspended and from the typical users, we do not have yet an understanding of the behavior of the users who get suspended without other users flagging their abusive content. An initial analysis of these users—suspended but not flagged—shows the following particularities when compared to the fair users (all users, independent of their deviance status, who are not suspended).

First, they are followed by and follow significantly fewer other users. Figures 9 (a) and (b) show the distributions of indegree and outdegree of never-flagged-suspended users compared to those of fair users. Not only these users have smaller social circles, but they also have lower activity levels, as shown in Figure 9 (c). Of course, these results could be correlated: low activity may mean low engagement in the social platform. These results may also suggest that (some of) these users join the platform for particular objectives that are orthogonal to the platform purpose, such as spamming. More importantly, however, these results suggest directions that we present in the following.

6. MEMBERS OF THE NETWORK

We investigate how the social network defined by the follower-followee relationships impacts user activities and behaviors in *YA*. Our final goal is to understand how to separate fair users from users who should be suspended even in the absence of flags. We learn that users close in the *FF* network not only help each others by answering questions, but also monitor each other's behavior by reporting flags (Section 6.1). Thus, the social network allows users to implicitly coordinate their behavior so much so that users who are socially close exhibit not only similar behavior, but also a similar deviation from the typical behavior (Section 6.2).

6.1 Out of Sight, Out of Mind

We expect that users receive more answers from users that are close in the social network. To verify this intuition, we randomly selected 7M answers such that both parties of the dialogue (the user who posted the question and the user who answered it) are in the social network, and measured the social distances between the two users. For a user u and a social distance h, the probability of receiving an answer from followers at distance h is the following:

$$p_h = \frac{\#\ \text{of } u\text{'s followers at distance } h \text{ who answered } u\text{'s questions}}{\#\ \text{of } u\text{'s followers at distance } h} \tag{4}$$

Figure 10 plots the geometric average of all these probabilities at a given distance as a function of social distance. The figure confirms that the probability of receiving answers from h-hop followers decreases with social distance.

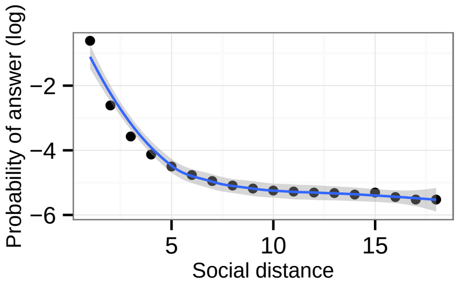


Figure 10: Probability of getting answers from h-hop followers. 95% confidence interval area is shown.

Therefore, the *FF* network channels user attention, likely via its newsfeeds feature that sends updates to followers

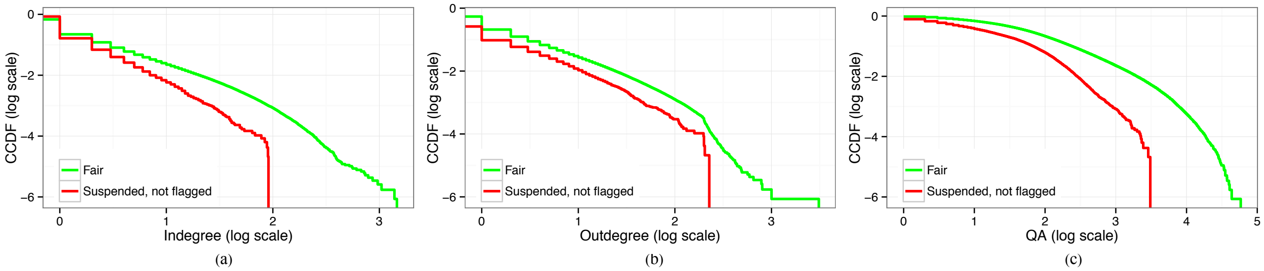


Figure 9: Distributions of (a) indegree; (b) outdegree and (c) number of questions and answers (QA) of never flagged suspended users and fair users. For outdegree: $D = 0.28$ and $Z = 27.40$, $p < 0.001$. For indegree: $D = 0.17$ and $Z = 15.86$, $p < 0.001$. For activity: $D = 0.30$ and $Z = 40.30$, $p < 0.001$.

on the questions posted by the user. Does the same phenomenon hold true for abuse reports?

To answer this question we investigate both networks: along with the *FF* which is an explicit network, we also investigate the activity network (AN), which connects users based on their direct interactions question-answer. For each (reporter, reportee) pair in the editorially-curated abuse reports, we calculated the shortest path distance between them in the social network and the activity network. We compare our results with a null model that randomly assigns the abuse reports in our sample dataset to users in the two networks.

Figure 11 shows the percentage of abuse reports users receive from close distances (up to 8 hops) for both (social and random) cases. About 75% of the reports that users receive are from reporters located within 5 social hops in the *FF* network. However, when reports are distributed randomly, about 9% are from within 5 social hops and very few from within 3 social hops.

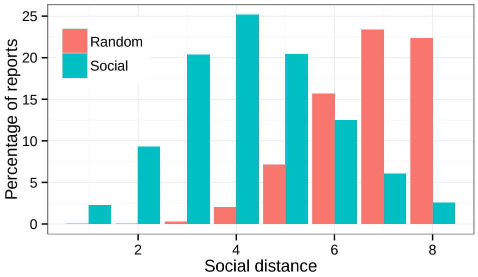


Figure 11: Percentage of the abuse reports received by users from different distances in the social network, for the observed case and a random case.

When comparing the percentage of abuse reports users receive with respect to distance in the AN (Figure 12), we notice that 94% of reports come from users within the first 3 hops, which is significantly higher than the social network (about 32%). We believe this is due to the high density of AN: most of the nodes are reachable from others within a few hops. However, even in this denser network, the null model has only about 10% of reports applied from within 3 hops.

To further quantify this phenomenon, we calculate the probability of being correctly flagged by users located at different network distances in the social and the activity network. For a user u and a social distance h, the probability of being flagged by followers at distance h is the following:

$$p_h = \frac{\#\text{ of } u\text{'s followers at distance } h \text{ who flagged } u}{\#\text{ of } u\text{'s followers at distance } h} \quad (5)$$

Figure 13 plots the geometric average of all probabilities at a given distance against the social distance for both networks. As expected, the probability decreases with social distances in both the social and the activity network. The plot shows that users are likely to receive flags from others close to them in terms of social relationships and interactions.

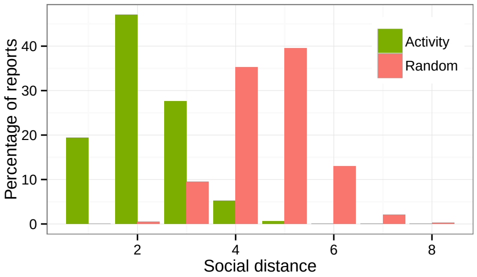


Figure 12: Percentage of the abuse reports received by users from different distances in the activity network, for the observed case and a random case.

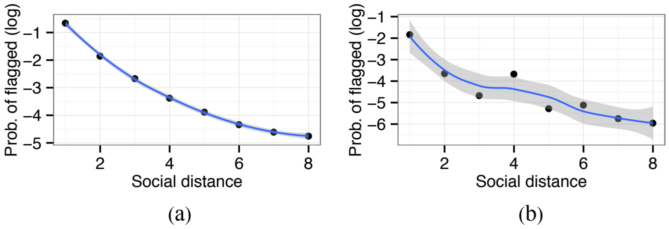


Figure 13: Probability of being flagged by h-hop followers (a) social network; (b) activity network. 95% confidence interval area is shown.

These results confirm that the abuse reporting behavior is dominated by social relationships and interactions: users are reported for content abuse more from their close social or activity neighborhoods than from distant users. The underlying reason is likely content exposure: a user's contents

Table 3: Assortativity coefficient r for deviance scores in the *YA* network. r values are also shown for other social networks from [21].

Yahoo! Answers	Other Social Networks
Question deviance $r = +0.11$	Mathematics coauthorship $r = +0.120$
Answer deviance $r = +0.13$	Biology coauthorship $r = +0.127$

(questions/answers) are disseminated to nearby followers, thus they get higher exposure to that content compared to more distant users in the social graph. Similarly, users who interact frequently with a user are more probable to view her contents and to report the inappropriate ones.

6.2 Birds of a Feather Flock Together

Similarity fosters connection– a principle commonly known as homophily, coined by sociologists in the 1950s. Homophily is our inexorable tendency to link up with other individuals similar to us [19]. We verify in this section whether homophily is also present in terms of deviance–that is, if deviant users tend to be close to each other in the social network.

One way to conclude about the homophily of a network is to compute the attribute assortativity of the network [22]. The assortativity coefficient is a measure of the likelihood for nodes with similar attributes to connect to each others. The assortativity coefficient ranges between -1 and 1; a positive assortativity means that nodes tend to connect to nodes of similar attribute value, while a negative assortativity means that nodes are likely to connect to nodes with very different attribute value from their own. If a network has positive assortativity coefficient, then it is often called assortative mixed by the attribute, otherwise called disassortative mixed.

In this work, we used question and answer-based deviant scores. We considered each of the scores as an attribute and calculated the assortativity coefficient r based on [21] for each type of deviance. The assortativity coefficients r are shown in Table 3 and are positive.

In [21], Newman studied a wide variety of networks and concluded that social networks are often assortatively mixed (Table 3), but that technological and biological networks (e.g., World Wide Web $r = -0.067$, software dependencies $r = -0.016$, protein interactions $r = -0.156$) tend to be disassortative. Comparing them quantitatively with the assortativity coefficients of the *YA* network, we conclude that the *YA* network is assortatively mixed in terms of deviance. So, users having contacts with (low)high deviance scores will also have (lower)higher deviance scores.

We next measure how similar the deviance scores of a user's contacts are with the user's, and how this similarity varies over longer social distances. For this, we randomly sampled 100,000 users from the social network for each social distance ranking from 1 hop to 4 hops.

Let U_h be the set of all the users (100,000) selected for the social distance h. We calculated the probability that user u's h-hop contacts (with $u \in U_h$) will have the same deviance score as:

$$p_u = \frac{\#\text{ of } u\text{'s followers at distance } h \text{ with same deviance score}}{\#\text{ of } u\text{'s followers at distance } h} \tag{6}$$

Rather than computing the exact similarity between a user and her follower's deviance scores, we focused on whether their difference is small enough to be dubbed as the same. We considered two users' deviance scores are the same if their corresponding deviance score difference is less than a "similarity delta". More specifically, u will have *about the same* deviance score with user s located at distance h if:

$$|deviance_u - deviance_s| < \delta \tag{7}$$

The same technique was used for both types of deviance scores. We experimented with two values for δ equal to one or two standard deviations of the distribution of deviance scores in the network. We report P_h, the geometric average of all p_u probabilities computed in each hop h:

$$P_h = \Big(\prod_{u \in U_h} p_u \Big)^{\frac{1}{|U_h|}} \tag{8}$$

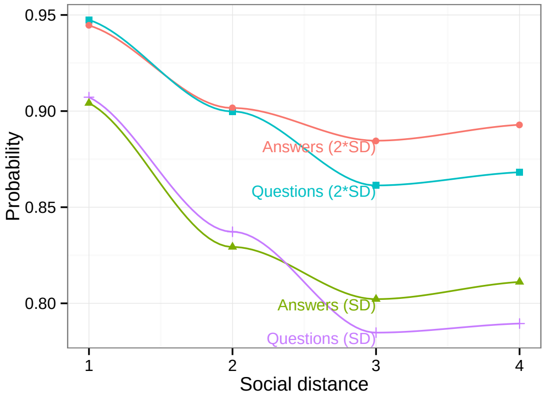


Figure 14: Probability that a h-hop follower has the same deviant score to the user for $\delta = \sigma$ and $\delta = 2\sigma$. SD: standard deviation.

Figure 14 shows the probability plots for both types of deviance, keeping similarity δ equal to one or two standard deviations. Although different values of the δ, the shapes of the figures are almost the same: the probability decreases gradually with the social distance.

7. SUSPENDED USER PREDICTION

Based on our previous analysis, we extract various types of features that we use to build predictive models. We formulate the prediction task as a classification problem with two classes of users: fair and suspended. Next, we describe the features used and the classifiers tested, and demonstrate that we are able to automatically detect fair from suspended users on Yahoo Answers with an overall high accuracy.

7.1 Features for Classification

Our predictive model has 29 features that are based on users' activities and engagements e.g., *social*, *activity*, *accomplishment*, *flag* and *deviance*. Table 4 shows the different categories of features used for the classification. *Social* features are based on the social network of the users, where *Activity* features are based on community contributions in the form of questions and answers. *Accomplishment* features acknowledge the quality of user contribution (e.g., points, best answers). *Flag* summarizes the flags of a user (both received and reported). *Deviance Score* features are the scores that we have got based on users' flags and activities. Finally, *Deviance Homophily* represents the homophilous behavior with respect to deviance. Although most of the features are self-explanatory, below we clarify the ones which may not be.

Reciprocity. Reciprocity measures the tendency of a pair of nodes to form mutual connections between each other [12]. Reciprocity is defined as follows:

$$r = \frac{L}{L^*}$$

where L is number of edges pointing in both directions and L^* is the total number of edges. $r = 1$ holds for a network in which all links are bidirectional (purely bidirectional network), while a purely unidirectional network has $r = 0$.

Status. Status is defined as follows:

$$Status = \frac{\#followers}{\#followees}$$

Thumbs. Thumbs is the difference between the number of up-votes and the number of down-votes a user receives for all her answers.

Award Ratings: Sum of the ratings a user receives for her best answers.

Altruistic scores: Difference between a user's contribution and his takeaway from the community. For altruistic scores, we consider *YA*'s point system, which awards two points for an answer, 10 points for a best answer, and penalizes five points for a question:

$$\begin{aligned} \text{Altruistic scores}_u &= f(contribution) - f(takeaway) \\ &= 2.0 * A_u + 10.0 * BA_u - 5.0 * Q_u \end{aligned} \tag{9}$$

where Q_u is the number of questions posted by u, A_u is the number of answers posted by u, and BA_u is the number of best answers posted by u.

7.2 Experimental Setup and Classification

In our dataset, the percentage of fair users (about 91%) are high compared to the suspended users (about 9%). This leads to an unbalanced dataset. Various approaches have been proposed in the machine learning literature to fix the unbalanced dataset. We use ROSE [20] algorithm to create a balanced dataset from the unbalanced one. ROSE creates balanced samples by random over-sampling minority examples, under-sampling majority examples or by combining over and under-sampling. Our prediction dataset has 250K users with 60-40% training–testing split. Using the under and over sampling technique of ROSE, we sample 150K users (fair and suspended each class has 75K users) to train the classifier. The testing set has 100K users, who are not present in the training dataset. They are drawn randomly and fair vs. suspended ratio in the testing dataset is the same as the original YA dataset.

We have used various classification algorithms, including Naive Bayes, K-Nearest Neighbors (KNN), Boosted Logistic Regression, and Stochastic Gradient Boosted Trees (SGBT) and found that the SGBT shows the best performance. SGBT offers a prediction model in the form of an ensemble of weak prediction models [11]. Table 5 shows a summary of our experimental setup. First, we use individual feature sets to investigate how successful one feature set is by itself only, and finally used all features for prediction. For evaluation, we measure widely used metrics in classification problems: Accuracy, Precision, Recall and F1-score.

Table 4: Different categories of features used for fair vs. suspended user prediction.

Category	Number	Features
Social	6	Indegree Outdegree Status Reciprocity Reciprocated networks degree Reciprocated networks CC
Activity	4	#Questions #Answers #Flagged Questions #Flagged Answers
Accomplishment	5	Points #Best Answers Award Ratings Thumbs Altruistic scores
Flag	8	#Question Flag Received #Question Flag Received Valid #Question Flag Reported #Question Flag Reported Valid #Answer Flag Received #Answer Flag Received Valid #Answer Flag Reported #Answer Flag Reported Valid
Deviance Score	2	Question deviance score Answer deviance score
Deviance Homophily	4	Followers' question deviance score Followers' answer deviance score Followees' question deviance score Followees' answer deviance score

Table 5: Details of experimental setup.

Dataset	250,000 users
Class Balancing Alg.	Random Over-Sampling Examples (ROSE)
Classifiers	Stochastic Gradient Boosted Trees (SGBT) Naive Bayes, Boosted Logistic Regression K-Nearest Neighbors (KNN) Support Vector Machines RDF
Feature Sets	Social, Activity, Accomplishment Flag, Deviance Homophily, All features
Train-Test Split	150K users training, 100K users testing
Cross Validation	10-folds, repeated 10 times
Performance	Accuracy, precision, recall, F1 score

7.3 Classification Results and Evaluation

Figure 15 shows the performance (accuracy, precision, recall and F1 score) of the models trained with different subsets of features using the Stochastic Gradient Boosted Trees (SGBT) classifier. We observe that each feature set has a positive effect on the performance of the classifier across all performance metrics. This suggests that all our feature sets are important for prediction. Particularly, accomplishment, deviance, flags and activity features individually can predict more than 70% accuracy with good precision, recall and F1 score. However, when all the features are used for classification, the performance metrics yielded best results.

The performance results of various classifiers while using all features are shown in Table 6. The SGBT classifier outperforms other classifiers in all performance metrics. It achieves 82.61% accuracy in classifying fair vs. suspended users with a high precision (96.94) and recall (83.52). The

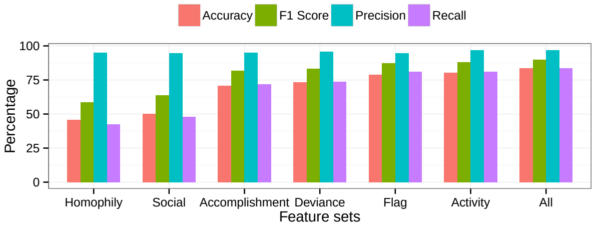


Figure 15: Performance of the SGBT while classifying fair and suspended users. Four performance measures are shown: Accuracy, Precision, Recall and F1 score.

confusion matrix of the classifier is shown in Table 7. The matrix shows that the SGBT classifier is able to correctly classify 83.52% of fair users and 73.39% of suspended users.

Table 6: Performance of various classifiers while using all features.

Classifier Name	Accuracy	Precision	Recall	F1 Score
Naive Bayes	47.21	96.93	43.34	59.89
Boosted Logistic Regression	71.61	96.62	71.28	82.03
KNN	73.81	96.41	73.97	83.71
SVM-RDF	75.92	95.62	77.06	85.34
SGBT	82.61	96.94	83.52	89.73

Table 7: Confusion matrix for SGBT classifier.

		Actual	
		Fair	Suspended
Predicted	Fair	**83.52**%	26.60%
	Suspended	16.47%	**73.39**%

Figure 16 shows the most important features (top 15) in classification of fair vs. suspended users. The number of flagged content and deviance scores are the best predictors of fair and suspended users. We can also observe at least one feature from all feature sets are within the top fifteen features. However, only activity and deviance score features sets have all the features within the top fifteen features.

8. SUMMARY AND DISCUSSIONS

This paper is an investigation of the flagging of inappropriate content in a popular and mature community Q&A, Yahoo Answers. Based on a year worth of activity records that included about 10 million flags in a population of about 1.5 million active users, our analysis revealed the following:

The use of flags is overwhelmingly correct, as shown by the large percentage of flags validated by human monitors. This is an important learning for crowd sourcing, as it shows for the first time (to the best of our knowledge) that crowd sourced monitoring of content functions well in CQA platforms. Moreover, although there are no explicit incentives (e.g., points) for flagging inappropriate content, users take the time to curate their environment. In fact, 46% of the users reported at least one abuse report, with the top abuse reporters flagging tens of thousands posts.

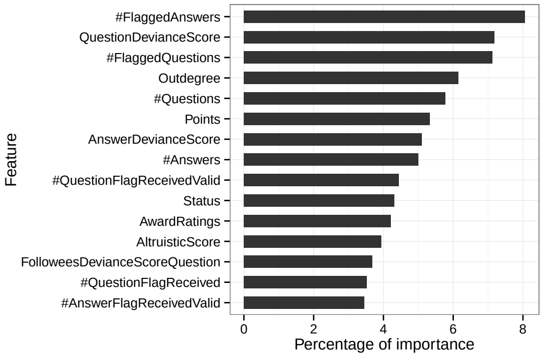


Figure 16: Relative importance of top 15 features in classifying fair and suspended users.

Second, we discover that many users have collected a large number of flags, yet their presence is not deemed toxic to the community. Even more, their contributions are engaging, which is certainly a benefit to the platform: the questions asked by the users who deviate from the norm in terms of number of flags received for their postings receive many more answers and from many more users than the questions posted by ordinary users or by users who later had their accounts suspended. More content-based analysis is needed to understand how the deviant users engage the community. We posit that they might ask conversational questions rather than informative questions, as this behavior is shown to increase community engagement.

Third, we showed the importance of the follower-followee social network for channeling attention and producing answers to question. Less expected, perhaps, is the fact that this network also channels the attention of flaggers: we showed that users in close social proximity are more likely to flag inappropriate content than distant users. Social neighborhoods, thus, tend to maintain their environment clean.

Fourth, a significant problem in *YA* is posed by the users who manage to avoid flagging, possibly by remaining at the outskirts of the social network. This relative isolation in terms of followers and in terms of interactions probably allows such users to remain invisible. They are likely caught by automatic spam-detection-like mechanisms and by paid human operators. Our empirical investigations show that classifiers that use activity-based features and social network-based features can successfully identify fair and suspended (40% of them are not flagged) users with an accuracy as high as 83%.

This work leads to various promising directions for future work. Understanding what makes deviant users engaging can be helpful in designing strategies potentially applicable to a variety of communities. Quantifying the equivalent behavior in terms of content abuse reporting and in terms of bad users on different online platforms can help understand the relative importance of different features for the success of the platform. And finally, characterizing the pro-social users who report abusive content in terms of both activity and social network centrality characteristics may help identify such potential volunteers and incentivize them appropriately.

9. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International*

Conference on Web Search and Data Mining, WSDM '08, pages 183–194, New York, NY, USA, 2008. ACM.

[2] Y. Answers. Yahoo answers community guidelines. `http://answers.yahoo.com/info/community_guidelines`, 2013.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[4] J. Bian, Y. Liu, E. Agichtein, and H. Zha. Finding the right facts in the crowd: Factoid question answering over social media. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 467–476, New York, NY, USA, 2008. ACM.

[5] J. Blackburn, R. Simha, N. Kourtellis, X. Zuo, M. Ripeanu, J. Skvoretz, and A. Iamnitchi. Branded with a scarlet "c": Cheaters in a gaming social network. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 81–90, New York, NY, USA, 2012. ACM.

[6] J. P. Davis. The experience of 'bad' behavior in online social spaces: A survey of online users. *Social Computing Group, Microsoft Research*, 2002.

[7] D. Dearman and K. N. Truong. Why users of yahoo!: Answers do not answer questions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 329–332, New York, NY, USA, 2010. ACM.

[8] J. D. Douglas and F. C. Waksler. *The sociology of deviance: An introduction.* Little, Brown Boston, MA, 1982.

[9] G. Dror, Y. Maarek, and I. Szpektor. Will my question be answered? predicting "question answerability" in community question-answering sites. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 499–514. Springer, 2013.

[10] P. D. Dunlop and K. Lee. Workplace deviance, organizational citizenship behavior, and business unit performance: The bad apples do spoil the whole barrel. *Journal of Organizational Behavior*, 25(1):67–80, 2004.

[11] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[12] D. Garlaschelli and M. I. Loffredo. Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93(26):268701, 2004.

[13] F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 759–768. ACM, 2009.

[14] F. M. Harper, J. Weinberg, J. Logie, and J. A. Konstan. Question types in social q&a sites. *First Monday*, 15(7), 2010.

[15] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 228–235, New York, NY, USA, 2006. ACM.

[16] B. Li, T. Jin, M. R. Lyu, I. King, and B. Mak. Analyzing and predicting question quality in community question answering services. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 775–782, New York, NY, USA, 2012. ACM.

[17] Q. Liu, E. Agichtein, G. Dror, Y. Maarek, and I. Szpektor. When web search fails, searchers become askers: Understanding the transition. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 801–810, New York, NY, USA, 2012. ACM.

[18] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann. Design lessons from the fastest q&a site in the west. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2857–2866, New York, NY, USA, 2011. ACM.

[19] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[20] G. Menardi and N. Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122, 2014.

[21] M. E. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[22] M. E. J. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[23] K. Panovich, R. Miller, and D. Karger. Tie strength in question & answer on social network sites. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1057–1066, New York, NY, USA, 2012. ACM.

[24] D. J. Phillips. Defending the boundaries: Identifying and countering threats in a usenet newsgroup. *The information society*, 12(1):39–62, 1996.

[25] M. Qu, G. Qiu, X. He, C. Zhang, H. Wu, J. Bu, and C. Chen. Probabilistic question recommendation for question answering communities. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1229–1230, New York, NY, USA, 2009. ACM.

[26] M. Richardson and R. W. White. Supporting synchronous social q&a throughout the question lifecycle. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 755–764, New York, NY, USA, 2011. ACM.

[27] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community qa. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 411–418, New York, NY, USA, 2010. ACM.

[28] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor. Learning from the past: Answering new questions with past answers. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 759–768, New York, NY, USA, 2012. ACM.

[29] J. R. SULER and W. L. Phillips. The bad boys of cyberspace: Deviant behavior in a multimedia chat community. *CyberPsychology & Behavior*,

1(3):275–294, 1998.

[30] K. Sun, Y. Cao, X. Song, Y.-I. Song, X. Wang, and C.-Y. Lin. Learning to recommend questions based on user ratings. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 751–758, New York, NY, USA, 2009. ACM.

[31] I. Szpektor, Y. Maarek, and D. Pelleg. When relevance is not enough: Promoting diversity and freshness in personalized question recommendation. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1249–1260, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[32] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the social crowd: An analysis of quora. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1341–1352, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[33] I. Weber, A. Ukkonen, and A. Gionis. Answers, not links: Extracting tips from yahoo! answers to address how-to web queries. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 613–622, New York, NY, USA, 2012. ACM.

[34] J. M. Wellen and M. Neale. Deviance, self-typicality, and group cohesion the corrosive effects of the bad apples on the barrel. *Small Group Research*, 37(2):165–186, 2006.