Model for Voter Scoring and Best Answer Selection in Community Q&A Services

Chong Tong Lee[*1], Eduarda Mendes Rodrigues[†2], Gabriella Kazai[†3], Nataša Milić-Frayling[†4], Aleksandar Ignjatović[*5]

[*]School of Computer Science and Engineering
University of New South Wales
Sydney, 2052, Australia
[1]ctlee@cse.unsw.edu.au, [5]ignjat@cse.unsw.edu.au

[†]Microsoft Research
7 JJ Thomson Avenue
Cambridge, CB3 0FB, UK
[2]eduardamr@acm.org, [3]gabkaz@microsoft.com,
[4]natasamf@microsoft.com

***Abstract*—Community Question Answering (cQA) services, such as Yahoo! Answers and MSN QnA, facilitate knowledge sharing through question answering by an online community of users. These services include incentive mechanisms to entice participation and self-regulate the quality of the content contributed by the users. In order to encourage quality contributions, community members are asked to nominate the 'best' among the answers provided to a question. The service then awards extra points to the author who provided the winning answer and to the voters who cast their vote for that answer. The best answers are typically selected by plurality voting, a scheme that is simple, yet vulnerable to random voting and collusion. We propose a weighted voting method that incorporates information about the voters' behavior. It assigns a score to each voter that captures the level of agreement with other voters. It uses the voter scores to aggregate the votes and determine the best answer. The mathematical formulation leads to the application of the Brouwer Fixed Point Theorem which guarantees the existence of a voter scoring function that satisfies the starting axiom. We demonstrate the robustness of our approach through simulations and analysis of real cQA service data.**

Keywords-community question answering; weighted voting; FPS method; voter score; vote spam; fixed point theorem

I. INTRODUCTION

Community Question Answering (cQA) services, such as MSN QnA and Yahoo! Answers leverage user interactions and user generated content in order to satisfy the information needs of the community members. They complement traditional search and information portals with support for sharing knowledge and delivering prompt responses to users' questions. Users exchange information by asking and answering each others' questions.

Considering the breadth and diversity of the cQA communities, it is natural to question the quality of the exchanged content. Su et al. [14], for example, found that the quality of answers varies significantly, while Harper et al. [8] showed that the quality of answers can compare and even surpass that of experts and library reference services, despite the different levels of users' expertise. This aspect of cQA gave rise to research that is focused specifically on the characterization of answers quality [1, 2, 10] and the users' authority [6, 12]. Our work complements these efforts by focusing on the mechanisms by which cQA services aim to promote quality content. More precisely, we study the approaches adopted by MSN QnA and Yahoo! Answers to provide incentives for community members to vote for *best answers*. In order to keep the incentive scheme transparent, cQA services typically use simple *plurality voting* to select best answers and apply a *winner-takes-all* approach to award additional points to 'successful' voters and answerers. Similarly to [3], we recognize that such incentive models can easily be misused through increased and superficial voting for the purpose of personal or social gain. This could ultimately lead to the failure of the community voting system as the means for identifying quality content and rewarding users for quality contributions.

In this paper we introduce a method that assigns a score to each voter which reflects the level of the voter's agreement with others when voting for the best answer. We then derive a *weighted voting* scheme for selecting the best answer by which a vote assigned to an answer is weighted by the score of the voter.

Our research contribution is a new axiomatically founded mathematical model for selecting best answers and a family of functions for voter scoring that are compatible with the starting axioms. The mathematical formulation leads to the application of the *Brouwer Fixed Point Theorem* [7] to prove the existence of the desired voter scoring functions and arrive at an iterative computational procedure for calculating the voter and answer scores. Through simulation experiments we demonstrate the robustness of the model and, through the use of data from MSN QnA, we show its applicability to real service scenarios. These illustrate the advantages of our approach: the voter scores are intrinsically related to the user's overall voting consistency with co-voters across a set of questions and thus difficult to manipulate. As a result, it enables surfacing of quality answers as the best answers.

In the next section we define the problem and discuss related work, and in section III we present an axiomatic formulation of the mathematical model we consider. In section IV we provide results of our simulation experiments

and illustrate the application of our approach to an existing cQA service. We conclude by discussing the results and directions for further work.

II. BACKGROUND

A. Selection of Best Answers in cQA

In this section we describe the commonly adopted voting procedure by which the *best answers* are selected in cQA services. For a given question, the community of users provides a set of answers and, subsequently, engages in simple plurality voting to determine the best answer. A user can cast vote for only one answer among all the responses to a question. The votes are not revealed to the community until the voting period is over, at which point the best answer is declared by selecting the answer with the most votes. Throughout the paper we refer to this simple vote counting as the *VC method*.

The community members are expected to vote based on their true conviction about the quality of the answer. Any deviation from that voting principle can be considered as sub-optimal from the service point of view. However, the reasons for departing from that principle may vary:

Different best answer connotations. The community engages in answering a range of question types, from seeking factual information to posing puzzles and riddles to entertain the community. The notion of the best answer thus depends on the context and the nature of the question, ranging from the correctness and usefulness of the answer to its entertainment value [1]. The assessment criteria are not prescribed but evolve over time and are not necessarily uniform across users.

Social bias. Since the authors of the questions and answers are identifiable through their usernames, the assignment of votes may be influenced by social and personal ties developed through the cQA interaction, including the voter's perception, familiarity, and preferential treatment of community members [16].

Self-promotion. The service offers incentives to individuals to engage and contribute to the community. For that users are recognized and rewarded with a higher status and increased visibility in the community using a variety of mechanisms, like reputation points. For example, MSN QnA encourages users to answer questions by awarding them 5 points per answer and 20 points per best answer. Similarly, the voters gain 1 point for casting a vote and 4 points for voting for the winning best answer. Individuals' aspirations to excel in their social status can lead to behaviors that are overly focused on personal gain and adversely affect the quality of their contribution to the community.

B. Voting Issues

We are particularly concerned with voting behaviors that try to exploit the incentive mechanisms instigated by self-promotion. Such behaviors can potentially reach a level of subversion, e.g., through: (1) random voting by an individual in order to increase the number of points through voting, or (2) organized collusion by users who create an alliance and coordinate the voting in order to increase the likelihood of higher gains for each other. The latter is evidenced by the following question posed by a member of the MSN QnA community:

"*Microsoft or Apple? Feel free to argue and point out their good and bad points. Also feel free to rebut or debate on other people's standpoint. Best argument/ answer will get my friends' and my "best answer" reward.*"

In our work we are focusing on methods for aggregating votes and determining best answers that are resistant to these types of subversions.

C. Related Work

Most of the research in cQA has focused on methods to identify high quality content and predict best answers. For example, Agichtein et al. [2] combined the graph features of the social network with the content and usage features to train a classifier for determining answer quality. They have investigated a wide range of features (e.g., hubs and authority scores, user clicks, answer length), and shown that combining several different types of features leads to increased classification accuracy. In [4], Bian et al. presented a ranking framework for retrieving factual information that exploits patterns in the user interaction in order to retrieve high quality content in social media. Our approach differs in the sense that we aim to characterize and control the voting process that leads to the nomination of the best answers.

The need for considering the voting practice is needed because of random or coordinated voting which influence can be significant. Similar concerns were expressed in [2], concluding that the combination of content and usage features is likely to increase the classifier's robustness to spam. Indeed, with their approach it is not sufficient for an adversary to create content that deceives the classifier, but it must also simulate realistic user relationships or usage statistics. In [3], Bian et al. explicitly address the issue of malicious users who try to "play the system" by selectively promoting or demoting content for profit or fun. They developed a ranking framework for social media that uses machine learning to integrate user interactions and content relevance and create a scoring that is robust to common forms of vote spamming. Our work is similar in the aim to derive a voter and answer scoring that is robust to subversion, but we follow an analytical approach leading to a deterministic solution.

Aiming to characterize users in a cQA community, Jurczyk & Agichtein [12] apply link analysis (HITS) to the answer-to social network and demonstrate that the resulting authority score for a user is better correlated with the number of votes received by the user's answers than with the simple number of answers provided by the user. In our approach we characterize the user voting behavior without explicitly analyzing the answer-to graph.

The study of both the users and the user generated content by Adamic et al. [1] differentiates between types of

questions, which in turn reflect upon the criteria for selecting best answers. Their analysis involved 189 most active Yahoo! Answers topic categories, i.e., those categories with more than 1,000 posted questions in their data sample. They clustered the categories into 3 groups using k-means clustering on three primary features: thread length, content length, and asker-replier overlap. By inspecting the categories in each cluster, the authors suggest that the clusters correspond to three types of user activities: 1) discussion forums, 2) seeking advice and common sense expertise, and 3) asking factual information. More detailed analysis showed that the discussion categories normally contained questions with longer threads and users involved in these threads engaged in both posing and answering questions. In contrast, the factual categories had questions with shorter threads and users typically assumed only one of the roles, a questioner or a replier. By applying logistic regression with several simple features, including reply length, thread length, number of answers, and number of best answers by users, Adamic et al. were able to predict best answers with the average accuracy of 70%.

They also used an entropy measure to capture the degree of concentration or *focus* by individuals on a particular topic category. They expected that the lower user entropy, i.e., a focus on a specific topic of expertise, would yield higher proportion of best answers. However, they found no correlation between the total entropy of the user's activities across the categories and the overall percentage of best answers associated with the user. The authors attributed this to the overall diversity of question and answer types in cQA services, where only some types of questions require expertise. This was further confirmed by observing a positive correlation in case of specific types of categories, i.e., technical categories where factual information and domain knowledge are required. Furthermore, the authors noted that the voting scheme is not amenable to tracking the consistency in the quality of the content contributions. For example, the user's expertise is recognized only in instances when their answer is awarded the best answer status, even in situations when multiple equivalent answers may have been provided. The results of this research enabled us to contextualize our work on modeling the user's voting behavior in the cQA services.

A great body of related research also exists in the fields of peer-to-peer (P2P), multi-agent, or e-commerce systems, where various trust and reputation models have been explored to combat the effects of spam votes and adversarial attacks [11]. The solutions proposed there include the cluster filtering approach [5] or the use of majority opinion as in [13]. Our voter scores can be compared to the definition of reputation in these systems, but we do not rely on methods for trust propagation.

III. Voter Scoring and Best Answers

In this section we formulate the problem and define a mathematical model for the voter score.

A. Assumptions

Consider a set of questions Q and for each question $q \in Q$, the corresponding set of answers A_q. A group of community members V is engaged in voting for the best answers. Each member in V selects a set of questions to consider for voting and for each question casts a vote for only one of the answers that the question has received. We then aim to determine the best answer for a question q based on the resulting distribution of votes across all the answers in A_q. However, when aggregating the votes we want to take into account the overall voting behavior of each voter $v_i \in V$ expressed through a *voter score*. In particular, we consider the voter's level of agreement with other voters across the questions. We expect that such a metric will help with detecting anomalous voting practices.

The underlying premise is that the quality of answers is a reflection of the community opinion – there is no absolute judgment of quality or correctness that is external to the community. In the same spirit, the voting quality of an individual can be judged only relative to the community, e.g., based on how often that person voted for answers which were in the end declared the best answers.

Taking this one step further and assuming that the system has no information about the users' voting practice outside Q, both the voter scores and the best answer decisions are established simultaneously from the given distribution of votes. By making the best answer decision dependent on the voter scores (which, in turn, depend on how often the users' votes are associated with the best answers), we arrive at a circular definition that can be mathematically formulated as the *fixed point problem* as we show in the next section.

B. Voter Score

We start with the bootstrapping case, when the first question posed to the community is answered and the answers receive votes from the community members. Thus, the set of questions Q comprises a single question with multiple answers and multiple voters cast their votes to the answer of their choice. We first derive the formula for the voters' scores for this simple case and then propose its generalization for multiple questions.

Given a question q, consider two voters v_i and v_j who cast their votes, each making their choice of the best answer a_i and a_j, respectively. Consider all the voters V_q who vote on answers to q. They can be divided into three groups: those who selected a_i, those who selected a_j, and others who made a different selection of best answer from A_q. We would like the voter score to capture the agreement between voters and thus stipulate that the *relative scores of two voters are determined by the proportion of all the voters who made the same choice of the best answer:*

$$r_i : r_j = \mu(a_i) : \mu(a_j). \quad (1)$$

Here $\mu(a_t)$ denotes the total number of votes that the community of voters assigned to answer a_t chosen by v_t, and

r_ι is the score for voter v_ι. We can show that the following function satisfies the above property:

$$r_i = \sqrt{\frac{\sum_{\{v_\iota \in V_q\}} \{r_\iota : a_\iota = a_i\}}{\sum_{\{v_k \in V_q\}} (r_k)}}, \quad (2)$$

where the summation is over the group of voters V_q who voted for answers to q. In other words, the score r_i for the voter v_i is computed as the sum of scores of all voters who made the same choice of best answer for the given question as did voter v_i, normalized by the sum of scores of all voters who cast their votes across the answers A_q to q. Indeed, from formula (2) it is easy to see that:

a) $0 < r_i \leq 1$ for any choice of answers in A_q;

b) $r_i = r_j$ if $a_i = a_j$, for two voters v_i and v_j when Q contains one question.

By the definition of $\mu(a_i)$ and the observation in b), we can rewrite the sum in the numerator of (2) as:

$$r_i = \sqrt{\frac{\mu(a_i)\, r_i}{\sum_{\{v_k \in V_q\}} (r_k)}}. \quad (3)$$

Through a simple calculation,

$$(r_i)^2 = \frac{\mu(a_i) r_i}{\sum_{\{v_k \in V_q\}} (r_k)} \longrightarrow r_i = \frac{\mu(a_i)}{\sum_{\{v_k \in V_q\}} (r_k)}, \quad (4)$$

we show that the desired relationship (1) holds for any two voters.

Note that in case of multiple questions, it is suitable to replace the normalization factor in (2) by the sum of scores for all the voters, regardless of whether they are considering the same set of questions as voter v_i. Similarly, the top sum is undefined when there are no answers with a common vote and thus we can assign it the value of 0.

With that in mind, we write the formulation of the voter score as the arithmetic mean over $|Q|$, i.e., the number of questions in Q:

$$\left\{ r_i = \frac{1}{|Q|} \sum_{\{q \in Q\}} \sqrt{\frac{\sum_{\{v_\iota \in V_q\}} \{r_\iota : a_\iota = a_i\}}{\sum_{\{v_k \in V\}} (r_k)}} \right\}, \; 1 \leq i \leq |V|, \quad (5)$$

where $|V|$ designates the number of users voting on answers to questions in Q. Thus, the vector $\vec{r} = \langle r_1, \cdots, r_{|V|} \rangle$ of the voter scores is a fixed point for the function $\boldsymbol{F}$ when (5) is expressed as:

$$\boldsymbol{F}(\vec{r}) = \vec{r}. \quad (6)$$

1) Generalization

The relative voter scores need not be linear, as defined in (1). Super-linear scaling may suit a close competition among voters while sub-linear relation may help in a lopsided situation. Using a real parameter p and the modified function in (2):

$$r_i : r_j = (\mu(a_i))^p : (\mu(a_j))^p, \quad (7)$$

where $p > 1$ emphasizes the voter scores (super-linear relation), while $0 < p < 1$ de-emphasizes them (sub-linear relation). The respective solution is given by:

$$r_i = \left(\frac{\sum_{\{v_\iota \in V_q\}} \{r_\iota : a_\iota = a_i\}}{\sum_{\{v_k \in V\}} (r_k)} \right)^{1/\lambda}, \quad \lambda = \frac{p+1}{p}. \quad (8)$$

Furthermore, we can augment the class of functions to facilitate calculation of the voters' scores based on the voting activities over a fixed time period, modeling the *voting decay*. Indeed, similar to the approach in [9], we discount the voters' scores by a real parameter $\tau \geq 1$ using $t(q)$ as the closing time for the voting process associated with a question q:

$$\left\{ r_i = \frac{1}{\sum_{\{q \in Q\}} (\tau^{t(q)})} \sum_{\{q \in Q\}} \tau^{t(q)} \left(\frac{\sum_{\{v_\iota \in V_q\}} \{r_\iota : a_\iota = a_i\}}{\sum_{\{v_k \in V\}} (r_k)} \right)^{1/\lambda} \right\}, \; 1 \leq i \leq |V|. \quad (9)$$

In the next section we show that this extended class of functions meets the assumptions of the *Brouwer Fixed Point Theorem* and, therefore, there exists a voter score function $\vec{r}$ that satisfies (9).

2) Fixed Point Existence

In order to apply Brouwer Fixed Point Theorem [7] to the formulation in (6) for functions $\boldsymbol{F}$ defined in (5), (7), or (9), we must show that $\boldsymbol{F}$ is continuous, i.e., r_i is not arbitrarily close to zero. Consider:

$$\boldsymbol{F} : \{r_i : 1 \leq i \leq |V|\} \mapsto$$
$$\left\{ \frac{1}{\sum_{\{q \in Q\}} (\tau^{t(q)})} \sum_{\{q \in Q\}} \tau^{t(q)} \left(\frac{\sum_{\{v_\iota \in V_q\}} \{r_\iota : a_\iota = a_i\}}{\sum_{\{v_k \in V\}} (r_k)} \right)^{1/\lambda} : 1 \leq i \leq |V| \right\}. \quad (10)$$

We note that the expression on the right is largest, i.e., equal to 1, when all agents vote for the same answer. The smallest value is 0 if voter v_i did not vote for answers to q. However, voter v_i must have voted for at least one question in order to be part of the voters' community and have its score r_i used for the fixed point calculation. Hence, $0 < r_i \leq 1$.

Consider that v_i voted only once for an answer of the first question and no other voter agreed with v_i:

$$r_i = \frac{\min_{q \in Q} (\tau^{t(q)})}{\sum_{\{q \in Q\}} (\tau^{t(q)})} \left(\frac{r_i}{\sum_{\{v_k \in V\}} (r_k)} \right)^{1/\lambda} > \frac{\min_{q \in Q} (\tau^{t(q)})}{\sum_{\{q \in Q\}} (\tau^{t(q)})} \left(\frac{r_i}{|V|} \right)^{1/\lambda};$$
$$\therefore \; r_i > \frac{\min_{q \in Q} (\tau^{t(q)})}{\left(|V|\, |Q| \left(\max_{q \in Q} (\tau^{t(q)}) \right) \right)^{1/(\lambda - 1)}}. \quad (11)$$

Let's denote the expression on the right in (11) by ε. This is the lower bound of the voter score and the mapping can be written as $\boldsymbol{F} : [\varepsilon, 1]^{|V|} \mapsto [\varepsilon, 1]^{|V|}$. Hence a fixed point exists.

To compute the fixed point, we apply $\boldsymbol{F}$ iteratively, starting with the initial condition ($r_i = 1$, $i \le |V|$) until the difference between successive iterations is smaller than a threshold.

C. Answer Score

We now specify the Fixed Point Scoring (FPS) of individual answers based on the distribution of votes and the scores of voters who cast the votes. Given a question q and its corresponding set of answers $A_q = \{a_i : 1 \le i \le |A_q|\}$, where $|A_q|$ is the size of A_q, we calculate FPS as:

$$\mathrm{FPS}(a_i) = \sum_{\{v_t \in V_q\}} \{r_t : a_t = a_i\}. \quad (12)$$

For each question q we rank the answers according to their FPS and pronounce the highest scoring answer as the FPS *best answer*. This contrasts with the simple VC method, typically used by cQA services.

IV. Experiments

In the following sections we describe simulation experiments that reveal the strengths and limitations of the FPS approach to scoring voters and answers. We run simulations on synthetic data that comprises 1,000 questions and simulated user activities based on the properties of a same size sample taken from the MSN QnA service. We generate data distributions for: (1) the number of answers per question, (2) the number of votes per question and per individual answers, and (3) the frequency of voting by individual voters. Using such data as a foundation, we explore the robustness of the FPS method in two scenarios of user behaviors: random voting and coordinated voting.

A. MSN QnA Dataset

From the MSN QnA service we collected 488,760 questions, 1,330,819 corresponding answers (719,390 of which received votes from the community) and 1,599,994 votes. This content was contributed by 256,950 distinct users, 11.7% of which cast votes. Among the voters, a minority of 7.1% contributed 90% of all votes. We use a sample of most recent 1,000 questions to guide the synthetic data generation. By experimenting with a range of distribution models and parameters we settled on the following best-fit distributions.

Number of answers per questions. The number of answers per question with at least one vote, N_a, follows a geometric distribution with $p = 0.3$:

$$\Pr(x = N_a) = (1-p)^{N_a - 1} p. \quad (13)$$

Number of votes per question and answers. Further, we note that the number of votes k for a given question must be equal or larger than the number of answers N_a with at least one vote. Again, based on the sample inspection, we model the real vote distribution as a negative binomial distribution with $p = 0.3$, coincidently the same as for the distribution of answers per question:

$$\Pr(x = k - N_a) = \binom{k+N_a-1}{N_a-1} p^{N_a} (1-p)^k. \quad (14)$$

To characterize the distribution of votes across the set of answers to a question q, we consider the voting entropy. By definition, the voting entropy is maximized when votes are distributed uniformly across the set of answers and minimized when one of the answers receives all the votes. We found that the voting entropy increases with the total number of answers N_a, per question and use Zipf's law:

$$f(i) = \frac{i^{-s}}{\sum_{k=1}^{N_a} k^{-s}}, \quad (15)$$

to model the entropy increase, where $s = 1.5$ and i denotes the rank of an answer relative to other answers in A_q, based on the number of votes it received.

User voting activities. Uneven participation is a known property of large-scale online communities [15]. Services that rely upon a community of users to contribute content, ratings, votes, or similar, share a common trait: most of the contributions originate from a small percentage of users. cQA services are not an exception (e.g., see distribution of questions and answers per users in Yahoo! Answers [6]). The power law that characterizes uneven participation is effectively represented by the Zipf-Mandelbrot law in our discrete situation, with $q = 13$, $s = 1.8$:

$$f(i) = \frac{(i+q)^{-s}}{\sum_{k=1}^{N_v} (k+q)^{-s}}. \quad (16)$$

B. Experiment Design

In this section we present experiments with synthetic data which simulate two types of behaviors that affect information sharing and community building in cQA. We also discuss the application of the FPS method to real data from the MSN QnA service.

1) Simulation of Random Voting Behavior

The first scenario refers to a random assignment of votes by users across questions. This may be motivated by the cQA incentives for participation. Indeed, the cQA services encourage participation by giving out reward points and one way to increase participation is to vote frequently. However, for the services it would be undesirable if users simply cast votes without making an effort to assess the quality of the answers. In the extreme case, the users could randomly pick answers and vote for them, collecting points from rapid and high volume voting. We simulate random voting by adding to the original, Zipf's law distribution of votes across the answers A_q, another uniformly distributed set of votes. For a given percentage of random voters and a random selection of questions (e.g., 1% to 10% of the total questions), we

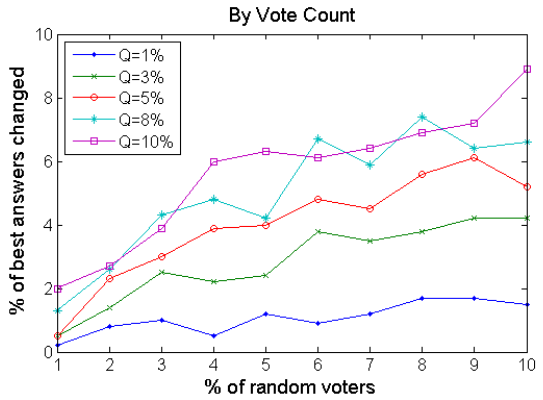

Figure 1. Random vote effect on the % of best answers changed by the VC method, when Q questions are affected by a given % of random voters.

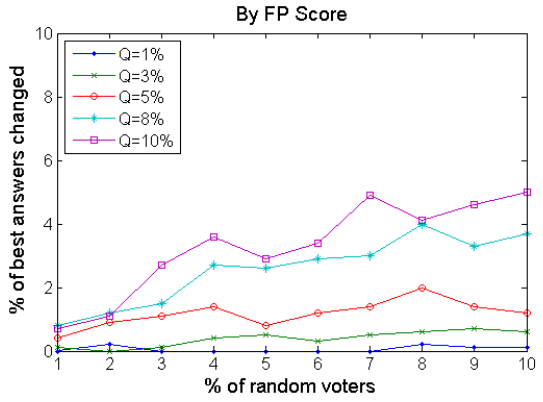

Figure 2. Random vote effect on the % of best answers changed by FPS method, when Q questions are affected by a given % of random voters.

assign a vote to a randomly selected answer in A_q, for each randomly selected question q.

The results show a higher level of robustness to random voting of the FPS method, in comparison to the VC method. With the introduction of random voting we observe the changes in the status of the previously chosen best answers. We note that the changes are more frequent when the best answers are nominated based on the VC method (Figure 1), than when they are selected using the FPS method (Figure 2). The FPS method is more robust because the random voters are given lower scores, as they cannot consistently assign their votes across questions. Thus, their votes cannot affect the relative ranking of answers in a major way.

2) *Voting in MSN QnA*

Generally, increased voting activity by individuals who wish to promote their standing in the community is not necessarily desirable for the service, since there is a greater chance of careless voting, similar to the random voting we experimented with. In many instances cQA services do not offer strong incentives for reliable voting: (1) the voting activity is not directly reflected in the user's reputation and (2) the reward for selecting the best answer is not significantly higher than otherwise, especially for users who vote a lot. This is in contrast with relatively high rewards for users who provide answers that are voted the best answer. In the next section we shall reflect on possible implications of such imbalance. Here we show that the FPS method has beneficial effects when applied to the data from a real cQA service, more specifically the MSN QnA data.

TABLE I. CHANGES IN BEST ANSWER SELECTION WITH THE FPS ANSWER SCORING.

Tag Sample	Total Answers	Best Answers Changed
Fun	21389	7050 (32.96%)
Technology	8478	2568 (30.3%)
Philosophy	9910	3799 (38.3%)

For this analysis we introduce a *voter's success rate* as the fraction of all the answers voted for by the user that turned out to be the best answers. For example, a success rate of 0.5 indicates that 50% of the time the voter chose answers that were subsequently declared the best for a given question. We can now observe the success rate of users in relation to the method applied to declare the best answer; in particular, the VC method compared to the FPS method. For this we consider questions tagged with three of the most popular tags from the MSN QnA community: 'Fun', 'Technology', and 'Philosophy'. For each tag, we segment the users based on their voting activity into three categories: 1) those who voted for less than 100 answers, 2) between 100 and 500 answers, and 3) more than 500 answers.

We calculate their voting success based on the VC and the FPS methods and compare the distribution of success rates across the voters (Figure 3). The histograms in the top row of Figure 3 refer to the VC method and show a great diversity in the users' voting success. The histograms in the bottom row correspond to the distribution of success rate for the FPS method and distinctively show that FPS rewards users who vote more actively. Combined with the robustness to random voting that we established in the previous section, this makes FPS a desirable method that can be used robustly with simple incentives that promote user participation. Table I shows the changes in the best answers statistics when switching to the FPS method.

3) *Simulation of Ballot Stuffing*

High rewards for the best answers may motivate a type of subversive behavior that is often referred to as 'ballot stuffing'. In this scenario, a subversion organizer asks a set of 'friends' to vote for the same answer he has already chosen so that the answer gets more votes and hence increasing its chance of becoming the best answer. We simulate this type of behavior by adding different numbers of colluding voters and varying the percentage of questions that the organizer has voted on.

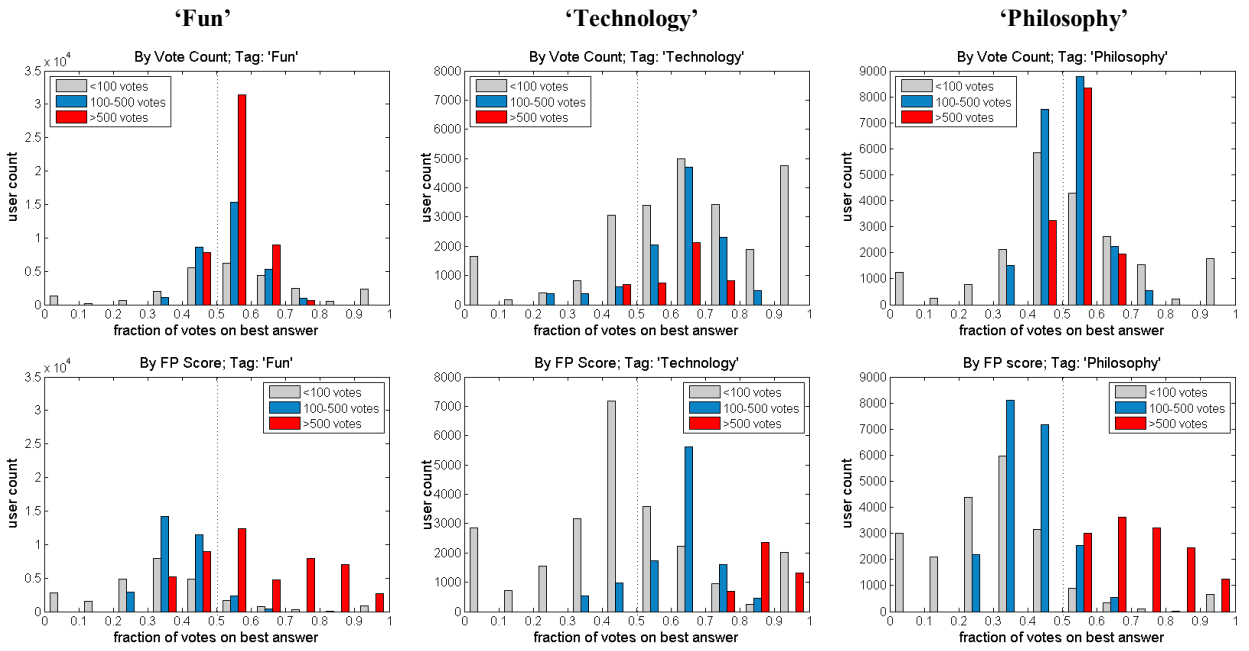

Figure 3. Distributions of users voting success in selecting the best answer, measured by the VC method (top plots) and by the FPS method (bottom plots), for questions tagged with 'Fun', 'Technology' and 'Philosophy'. Users are grouped into three classes that reflect their level of participation: (1) users with fewer than 100 votes (gray), (2) with between 100 and 500 votes (blue), (3) with more than 500 votes (red).

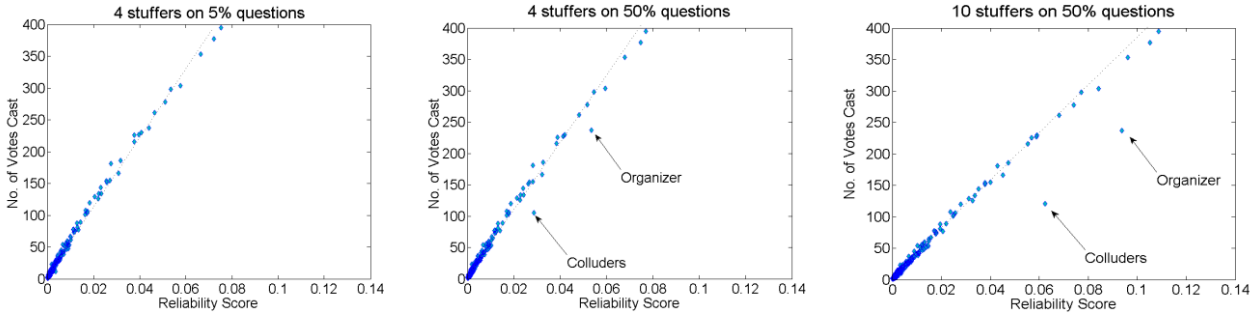

Figure 4.Voters' reliability scores vs. number of votes cast across the question threads, in 3 simulation scenarios: (1) 4 stuffers on 5% of questions, (2) 4 stuffers on 50% of questions, (3) 10 stuffers for 50% of questions.

Unlike random voting, collusion meets the fundamental principle of the FPS method, which relies on the level of agreement among users and their voting consistency across questions to award a higher voter score. Thus it appears that the method would not be able to detect the subversive behavior, but treat it in the same way as the genuine community agreement on the best answer choices and thus influence the best answer selection. However, when a large number of questions are involved, the exceptional ability of colluding voters to choose best answers more consistently than others makes them stand out in the overall distribution of voter scores (see Figure 4). They are observed as outliers. Hence, our method is useful for flagging anomalies and detecting users who may be colluding.

V. CONCLUSIONS

In our work we introduce a weighted voting method that is applicable to selecting best answers based on the votes that are cast by members of the cQA community. This method involves voter scoring which captures how often the voter's selection of best answers agrees with the choices of other voters. We start the model with an axiom stating that the relative reliability of two voters is proportional to their respective levels of agreement with other voters in the community. This axiom led to the formulation of the voter scores as the fixed point of a well behaved function. Thus we can prove the existence of a voter scoring function that satisfies the axioms and calculate its values through an

iterative computation process. The aggregate score for an answer is thus obtained as a weighted sum of the votes cast by individuals, with weights representing the voters' scores. We refer to this approach as the Fixed Point Scoring (FPS) method. Through simulation experiments we test the robustness of the FPS method for scoring answers and selecting the best answers. We show that the method is more robust to random voting than the typical vote counting method. Moreover, we demonstrate on a sample dataset from the MSN QnA service that the FPS approach rewards high level of engagement and can curb the effects of undesirable random voting.

The voters' scores also provide an additional metric for characterizing user behavior in real communities. Since the scores are sensitive to agreement and potential coordination of voting, ballot stuffing situations can be detected by looking at the outliers in the distributions of the scores over different levels of voting activity. Preliminary results are promising. We intend to explore a possible synergy between the FPS and other statistical methods such as correlation analysis of cQA interaction in order to detect small to medium scale coordinated voting behavior.

REFERENCES

[1] Adamic, L.A., Zhang, J., Bakshy, E., and Ackerman, M.S., "Knowledge Sharing and Yahoo Answers: Everyone Knows Something", In *Proc .of the 17th Int'l Conf. on World Wide Web (WWW '08)*, 2008, pp. 665-674.

[2] Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G., "Finding High-quality Content in Social Media", In *Proc. of the Int'l Conf. on Web Search and Web Data Mining (WSDM '08)*, 2008, pp. 183-193.

[3] Bian, J., Liu, Y., Agichtein, E., and Zha, H., "A Few Bad Votes Too Many?: Towards Robust Ranking in Social Media", In *Proc. of the 4th Int'l Workshop on Adversarial Information Retrieval on the Web (AIRWeb '08)*, 2008, pp. 53-60.

[4] Bian, J., Liu, Y., Agichtein, E., and Zha, H., "Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media", In *Proc .of the 17th Int'l Conf. on World Wide Web (WWW '08)*, 2008, pp. 467-476.

[5] Dellarocas, C., "Immunizing Online Reputation Reporting Systems against Unfair Ratings and Discriminatory Behavior", In *Proc. of the 2nd ACM Conf. on Electronic Commerce*, Minneapolis, MN, 2000, pp. 150-157.

[6] Gyongyi, Z., Koutrika, G., Pedersen, J., Garcia-Molina, H., "Questioning Yahoo! Answers", First Workshop on Question Answering on the Web, held at WWW, 2008.

[7] Griffel, D. H., *Applied Functional Analysis*, Dover Publications, 2002.

[8] Harper, F.M., Raban, D., Rafaeli, S., and Konstan, J. A., "Predictors of Answer Quality in Online Q&A Sites", In *Proc. of the 26th Annual SIGCHI Conf. on Human Factors in Computing Systems (CHI '08)*, 2008, pp. 865-874.

[9] Ignjatovic, A., Foo, N., and Lee, C.T., "An Analytic Approach to Reputation Ranking of Participants in Online Transactions", In *Proc. of IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI '08)*, 2008, pp. 587–590.

[10] Jeon, J., Croft, W., Lee, J., and Park, S., "A framework to Predict the Quality of Answers with Non-textual Features", In *Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '06)*, 2006, pp. 228-235.

[11] Jøsang, A., Ismail, R., and Boyd, C., A Survey of Trust and Reputation Systems for Online Service Provision. *Decision Support Systems*, 43(2), 2007, pp. 618-644.

[12] Jurczyk, P. and Agichtein, E. "Discovering Authorities in Question Answer Communities by Using Link Analysis", In *Proc. of the 16th ACM Conf. on Information and Knowledge Management (CIKM '07)*, 2007, pp. 919-922.

[13] Sen, S., and Sajja, N., "Robustness of Reputation-based Trust: Boolean Case", In *Proc. of Int'l Conf. on Autonomous Agents and Multi-Agents System (AAMAS '02)*, 2002, pp. 288-293.

[14] Su, Q., Pavlov, D., Chow, J., and Baker, W., "Internet-scale Collection of Human-reviewed Data", In *Proc. of the 16th Int'l Conf. on World Wide Web (WWW '07)*, 2007, pp. 231-240.

[15] Whittaker, S., Terveen, L., Hill, W., and Cherny, L., "The Dynamics of Mass Interaction", In *Proc. of the 1998 ACM Conf. on Computer-Supported Cooperative Work, (CSCW '98)*, 1998, pp. 257-264.

[16] Poston, R.S., "Using and Fixing Biased Rating Schemes", *Commun. ACM 51(9)*, 2008, pp. 105-109.