Quality Expectation-Variance Tradeoffs in Crowdsourcing Contests

Xi Alice Gao
Harvard SEAS
xagao@seas.harvard.edu

Yoram Bachrach
Microsoft Research
yobach@microsoft.com

Peter Key
Microsoft Research
peter.key@microsoft.com

Thore Graepel
Microsoft Research
thore.graepel@microsoft.com

Abstract

We examine designs for crowdsourcing contests, where participants compete for rewards given to superior solutions of a task. We theoretically analyze tradeoffs between the expectation and variance of the principal's utility (i.e. the best solution's quality), and empirically test our theoretical predictions using a controlled experiment on Amazon Mechanical Turk. Our evaluation method is also crowdsourcing based and relies on the peer prediction mechanism. Our theoretical analysis shows an expectation-variance tradeoff of the principal's utility in such contests through a Pareto efficient frontier. In particular, we show that the simple contest with 2 authors and the 2-pair contest have good theoretical properties. In contrast, our empirical results show that the 2-pair contest is the superior design among all designs tested, achieving the highest expectation and lowest variance of the principal's utility.

Introduction

Recently crowdsourcing has become a popular means to recruit individuals with diverse expertise and solve problems in a distributed fashion. Platforms such as Amazon Mechanical Turk (AMT, www.mturk.com) and Taskcn (www.taskcn.com) introduce both opportunities and challenges for firms seeking to get high quality results through these mediums in a timely fashion.

Crowdsourcing is a convenient solution for many tasks, such as categorizing images, transcribing audio clips, designing company logos, and writing articles on a particular topic. Consider a principal who recruits several participants (henceforth, authors) for a creative task such as writing a tourism ad for a specific destination. In order to obtain high quality solutions, the principal faces the challenge of providing the authors with appropriate incentives to exert effort when performing the task.

To address this incentive problem, many crowdsourcing platforms such as Taskcn and Topcoder (www.topcoder.com) are structured as contests, offering rewards to participants who provide solutions of superior quality. Exerting effort is costly for the participants but increases their chance of receiving the reward, so they must decide on how much effort to exert for the task without knowing the effort exerted by the other participants. This type of competition provides incentives to the participants to exert effort and provide high quality solutions. Crowdsourcing contests of this kind have been proposed and analyzed in the literature (Moldovanu and Sela 2001; DiPalantino and Vojnovic 2009; Archak and Sundararajan 2009; Chawla, Hartline, and Sivan 2011; Liu et al. 2011). These models predict the behavior of the contest participants by characterizing their equilibrium strategies.

In a mixed strategy Nash equilibrium of such a crowdsourcing contest, each participant's effort is a random variable drawn from an equilibrium distribution over the possible effort levels. *The principal's utility* is the quality of the best solution produced in the contest, which is also a random variable depending on the equilibrium strategies.

Existing analyses focus on computing the principal's expected utility under various designs, ignoring her risk of obtaining low quality solutions. For real-world applications, the principal desires not only a high expected utility but also low risk or *variance* in her utility. To address this weakness, we adopt a simplified model of crowdsourcing contests and characterize a trade-off between the expectation and variance of the principal's utility for different contest designs. Moreover, there is little *empirical* data from *controlled experiments* on the relation between the contest design and the principal's utility, although several papers examine contest designs already deployed (Archak 2010; Boudreau, Lacetera, and Lakhani 2011). Thus, we run controlled experiments on AMT to investigate the impact of the contest design on the principal's utility.

Further, theoretical analyses of crowdsourcing contests typically assume that the efforts exerted can be directly observed or uses simplified voting mechanisms for evaluating the qualities of the solutions. These approaches for quality evaluation are inadequate for several reasons. Many real-world tasks such as writing articles and designing company logos are *creative tasks*. For such tasks, quality evaluation is a highly subjective matter and a difficult task by itself, but is nonetheless required by an end-to-end framework for completing these tasks through crowdsourcing platforms. Ideally, the principal could also crowdsource the evaluation task, for example, by recruiting some voters to vote on a solution. Such a simple voting mechanism, however, may not provide appropriate incentives for the voters

to report their opinions honestly. To this end, we propose to use an existing mechanism, called *the peer prediction method*, to elicit honest subjective feedback from the participants (henceforth, critics) about the qualities of the solutions (Prelec 2004; Miller, Resnick, and Zeckhauser 2005; Witkowski and Parkes 2012). Such methods have attractive theoretical truthfulness guarantees. We use the term *author-critic framework* to refer to an end-to-end system, which incorporates the crowdsourcing contest for collecting solutions and the peer prediction method for quality evaluation. We implement this end-to-end system on AMT and empirically examine the principal's expected utility and risk under different contest designs through a controlled experiment, comparing our findings with the theoretical predictions.

Preliminaries and Related Work

Several game theoretic models have been proposed for analyzing the behavior of participants in contests by modeling these contests as all-pay auctions (Moldovanu and Sela 2001; DiPalantino and Vojnovic 2009; Archak and Sundararajan 2009; Chawla, Hartline, and Sivan 2011; Liu et al. 2011; Siegel 2009). As opposed to such analysis, we focus not only on the expectation of the principal's utility but also on its variance, and the trade-offs between the two.

Behavior in contest settings has also been studied empirically, mostly using real-world datasets based on deployed mechanisms already running "in the field" (Archak 2010; Boudreau, Lacetera, and Lakhani 2011), rather than through controlled experiments. In contrast, we verify our theoretical predictions through a controlled experiment, which involves completing and evaluating the solution to a creative task in a crowdsourcing environment. Also, empirical studies of all-pay auctions such as (Gneezy and Smorodinsky 2006) may shed some light on participant behavior in these contests.

Our experimental methodology builds on *peer prediction methods* for evaluating the quality of solutions to the creative task. Such methods were proposed to elicit honest subjective feedback from users about their experiences with products of unknown quality (Miller, Resnick, and Zeckhauser 2005). Since the true subjective opinion is neither observable nor verifiable, peer prediction methods reward participants based on the correlation between their reported opinions to induce a truthful equilibrium outcome. Among variants of the initial peer prediction method (Miller, Resnick, and Zeckhauser 2005; Prelec 2004; Jurca and Faltings 2005; 2007; 2009; Witkowski and Parkes 2012), we adopt the one proposed by Witkowski and Parkes (2012) (WP), which has several advantages such as not relying on the common knowledge assumption, requiring only a finite number of participants, and being individually rational.

We apply the WP method as follows. For each solution, we assume that a critic receives the h (high) signal if the quality of the solution exceeds her prior expectation and that she receives the l (low) signal otherwise. *Before* examining the solution, critic i is asked to report her belief of the expected quality of the solution, denoted by $p_1^i \in [0,1]$. Then, after reading the solution and deciding whether its quality exceeded her prior expectation or not (i.e. whether she received a h or l signal), critic i is asked to report her updated belief of the quality, denoted as $p_2^i \in [0,1]$. The second report p_2^i is a prediction of the likelihood that another random critic receives a h signal by reading the solution. Given the two belief reports, we infer that critic i received the h signal if $p_2^i > p_1^i$, and that she received the l signal if $p_2^i < p_1^i$. The payment of critic i depends on her two belief reports and on the signal received by another random critic $r(i)$ evaluating the same solution. If critic $r(i)$ received the h signal, then the payment of critic i is $\sum_{j=1,2}(2p_j^i - (p_j^i)^2)$. Otherwise, if critic $r(i)$ received the l signal, critic i is paid $\sum_{j=1,2}(1-(p_j^i)^2)$. The overall rating of a solution is the fraction of h signals from among all the inferred signals. Witkowski and Parkes (2012) showed that for the above payment rules, truthful reporting is a Bayesian Nash equilibrium, making this method an incentive compatible mechanism.

The Principal's Utility — Expectation-Variance Tradeoffs

We study how different contest designs affect the principal's utility. We use some simplifying assumptions to make our theoretical analysis feasible. First, in contrast to some contest models, we assume that all authors are of equal skill, and that the quality of the solution produced by each author is equal to the amount of effort she exerted to produce the solution (see (DiPalantino and Vojnovic 2009) for a model with different skill levels). We thus refer to the effort exerted for a solution and the quality of the solution interchangeably. We also assume that the principal has a black-box which can be used to accurately determine the relative qualities of solutions. For the empirical analysis we use the WP method (Witkowski and Parkes 2012) as an implementation of such a black-box.

Simple Contest

Consider the simple contest described as follows. A principal recruits n authors to produce solutions to a task in a simple contest and offers a single reward m for the solution of the highest quality. In this contest, author i chooses to exert a certain amount of effort e_i, which is measured in the same monetary units as the reward. Each author must exert a non-negative amount of effort, and no rational author would exert more effort than the reward m. The participant who exerted the most effort, i.e. $\arg\max_i e_i$, wins the contest, gets the reward m, and obtains a utility of $m - e_i$, whereas any other losing author j has a utility of $-e_j$. Since the exerted effort cannot be recovered, this crowdsourcing contest is essentially an "all-pay" auction. The principal who runs the contest is interested in the best possible solution to the task at hand, and chooses the best solution, so her utility is $\max_i e_i$. Note that this is different from an all-pay auction in which the utility of the auctioneer is the sum of the bids made by the participants.

For this simple contest, it is easy to see that a pure strategy Nash equilibrium does not exist. Thus, we characterize a symmetric mixed Nash equilibrium of this simple contest and analyze the principal's utility.

Theorem 1. *For the simple contest with n authors, there exists a symmetric Nash equilibrium where each author plays the same mixed strategy with full support over $[0, m]$ given by the cumulative distribution function (CDF):*

$$F(x) = \left(\frac{x}{m}\right)^{\frac{1}{n-1}} \tag{1}$$

Proof. For $x \in [0, m]$, let $\mathrm{F}(x)$ and $\mathrm{f}(x)$ denote the cumulative distribution function (CDF) and the probability density function (PDF) of the mixed strategy. The expected payoff for exerting effort $x \in [0, m]$ is $u(x) = \mathrm{F}(x)^{n-1}m - x$.

In a mixed strategy Nash equilibrium, the expected payoff of each author to exert any amount of effort $x \in [0, m]$ is the same. Thus, we have $u(x) = c, \forall x \in [0, m]$, for a constant c. Then applying the fact that $\mathrm{F}(m) = 1$ since $\mathrm{F}(x)$ is a valid cumulative distribution function, we can derive that $c = 0$.

Therefore, at this symmetric Nash equilibrium, each author's mixed strategy has the CDF:

$$\mathrm{F}(x) = \left(\frac{x}{m}\right)^{\frac{1}{n-1}}$$

□

At the symmetric equilibrium given in Theorem 1, the principal's utility is the best solution's quality, which is a random variable drawn from a distribution depending on the equilibrium mixed strategies. In Theorem 2, we derive expressions for the expectation and the variance of the principal's utility.

Theorem 2. *Under the symmetric equilibrium in Theorem 1, the expectation and the variance of the quality of the best solution are:*

$$E(X_1) = \frac{mn}{2n-1} \tag{2}$$

$$V(X_1) = \frac{m^2 n(n-1)^2}{(3n-2)(2n-1)^2} \tag{3}$$

Proof. Let X_1 be the random variable for principal's utility with a PDF $\mathrm{f}_{\max}(x)$ and a CDF $\mathrm{F}_{\max}(x)$. Then we can derive the expectation $E(X_1)$ and the variance $V(X_1)$ as follows:

$$\begin{aligned}
\mathrm{F}_{\max}(x) =& \mathrm{F}(x)^n = \left(\frac{x}{m}\right)^{\frac{n}{n-1}} \\
\mathrm{f}_{\max}(x) =& \frac{d}{dx}\mathrm{F}_{\max}(x) = \frac{n}{m(n-1)}\left(\frac{x}{m}\right)^{\frac{1}{n-1}} \\
\mathrm{E}(X_1) =& \int_0^m x\mathrm{f}_{\max}(x)dx = \frac{mn}{2n-1} \\
\mathrm{V}(X_1) =& \int_0^m (x - \mathrm{E}(X_1))^2 \mathrm{f}_{\max}(x)dx \\
=& \frac{m^2 n(n-1)^2}{(3n-2)(2n-1)^2}
\end{aligned}$$

□

Surprisingly, by equation (2), as the number of authors approaches infinity, the *expectation* of the principal's utility decreases and approaches $\frac{m}{2}$. By equation (3), the *variance* of the principal's utility, equivalently the principal's risk, increases and approaches $\frac{m^2}{12}$ as the number of authors approaches infinity.

Intuitively, as the number of authors increases, for a fixed mixed strategy, the probability for each author to win decreases. Therefore, at the symmetric equilibrium, the mixed strategy of each author becomes skewed, putting more probability mass on exerting less effort. The expectation decreases because the effect of the skewing outweighs the effect of having more authors.

For a simple crowdsourcing contest with n authors, our analysis shows that the optimal design, both in terms of maximizing expectation and minimizing variance, is recruiting 2 authors, achieving an expected utility of $\frac{2m}{3}$ for the principal. However, the variance $\frac{m^2}{18}$ is quite large, resulting in high risk for the principal who desires a good guarantee for the quality she would obtain. We now explore alternative designs that can potentially achieve a lower variance.

Alternative Contest Designs

Suppose the principal has a fixed budget of $m = 1$ wlog. The principal's objective is to design the contest so as to maximize the expectation and to minimize the variance of the principal's utility. We say that a design A Pareto dominates design B if design A offers both a higher expectation and a lower variance. We wish to find an efficient frontier on which improving either the expectation or the variance must require sacrificing the other. We visualize our results on a scatter plot, where each point represents a specific design; its X-value is the *expectation* of the principal's utility and its Y-value is the *variance*. In the plot, we seek points in the bottom right corner, with high expectation and low variance.

Design 1: Simple Contest with $1 budget: For all simple contests with budget $m = 1$, by Theorem 2, the pair contest ($n = 2$) achieves the highest expectation $\frac{2}{3}$ and the lowest variance $\frac{1}{18}$ of the principal's utility, giving a single point on the efficient frontier.

Inspired by the above analysis, we consider the following design of dividing a total of n authors into equal sized groups each containing a small number of authors. Having a small number of authors in each group induces the authors to put more probability on higher effort levels. However, the maximum possible effort for each author is also smaller since the reward for each group is smaller. Below we examine how this design will affect the expectation and the variance of the principal's utility.

Design 2: Contests of groups of k authors: Let the principal divide n authors into $\frac{n}{k}$ groups of size k (assuming $1 < k < n$ and n divides k) and offer a reward $\frac{k}{n}$ to the best solution in each group. This design essentially groups $\frac{n}{k}$ simple contests of k authors together while keeping a unit budget. By Theorem 1, at a symmetric equilibrium, each au-

thor plays a mixed strategy over $[0, \frac{k}{n}]$ with CDF:

$$\mathrm{F}(x) = \left(\frac{nx}{k}\right)^{\frac{1}{k-1}} \tag{4}$$

Denote by X_2 the random variable for the principal's utility. The expectation and the variance of X_2 are:

$$\mathrm{E}(X_2) = \frac{k}{n+k-1} = O\left(\frac{1}{n}\right)$$

$$\mathrm{V}(X_2) = \frac{k^2(k-1)^2}{n(n+k-1)^2(n+2k-2)} = O\left(\frac{1}{n^4}\right)$$

As $n \to \infty$, the expectation and variance approach 0 on the order of $\frac{1}{n}$ and $\frac{1}{n^4}$ respectively. Among all such contests, the contest with groups of 2 authors ($k = 2$) dominate the rest as they achieve both higher expectation and lower variance. Note that the pair contest ($n = 2$) and the 2-pair contest ($n = 4$) achieve relatively high expected utility for the principal, but neither of them Pareto dominates the other. We now focus on contests with 2 or 4 authors and characterize an efficient frontier by varying the reward rules.

Design 3: Varying rewards of 2 sub-contests: Consider recruiting 4 authors and running 2 contests A and B with 2 authors in each. Let contest A and B offer rewards of r and $1-r$ respectively. Without loss of generality, we assume $r \in [0.5, 1]$. Let X_3 denote the random variable of the principal's utility. The expectation and variance of X_3 are:

$$E(X_3) = \int_0^{1-r} \frac{4}{r^2(1-r)^2} x^4 dx + \int_{1-r}^{r} \frac{2}{r^2} x^2 dx$$

$$V(X_3) = \int_0^{1-r} (x - E(X_3))^2 \frac{4}{r^2(1-r)^2} x^3 dx + \int_{1-r}^{r} (x - E(X_3))^2 \frac{2}{r^2} x dx$$

Numerical calculations for this design show that both expectation and variance increase as r increases from 0.5 to 1. In Figure 1, the relationship follows a concave curve.

Design 4: Two-phase contest with 4 authors: Consider a contest with 2 pairs of authors in which the reward is determined in two phases. After every author submits an entry, in the first phase, each pair of authors competes for a reward of $\frac{r}{2}$. Then in the second phase, the best solution among all gets a reward of $1 - r$. Note that the quad contest (design 1 with $n = 4$) and the 2-pair contest are extreme cases of this design for $r = 0$ and $r = 1$ respectively. We find the symmetric equilibrium for this contest.

Theorem 3. *At a symmetric equilibrium of the two-phase contest, the CDF $F(x)$ of the mixed strategy of each author is the solution to the following equation*

$$(1-r)F(x)^3 + \frac{r}{2}F(x) = x \tag{5}$$

Proof. At a symmetric equilibrium of this game, we use $\mathrm{F}(x)$ to denote the CDF of each author's mixed strategy over $[0, 1-\frac{r}{2}]$. By writing out the expression of each author's utility function in terms of the effort exerted and applying the definition of a CDF, we can derive the implicit expression for $\mathrm{F}(x)$ given in equation 5. The derivation is similar to the one used in Theorem 1 and we omit the derivation here. □

For each $r \in [0, 1]$, we can derive an explicit expression for the CDF of the mixed strategy. For example, when $r = \frac{1}{2}$, the CDF of the mixed strategy is given below.

$$\mathrm{F}(x) = \sqrt[3]{x + \sqrt{x^2 + \frac{1}{216}}} - \sqrt[3]{-x + \sqrt{x^2 + \frac{1}{216}}} \tag{6}$$

This example shows that the explicit expression of $\mathrm{F}(x)$ is cumbersome to work with analytically. Thus, we resort to numerically computing the values of the variance and the expectation of the principal's utility. In the following formulas, let $r \in [0, 1]$ and let X_4 denote the random variable of the quality of the best solution.

$$E(X_4) = \int_0^{1-\frac{r}{2}} x \left(\frac{d}{dx}\mathrm{F}(x)^4\right) dx \tag{7}$$

$$V(X_4) = \int_0^{1-\frac{r}{2}} (x - E(X_4))^2 \left(\frac{d}{dx}\mathrm{F}(x)^4\right) dx \tag{8}$$

Numerically calculations show that the expectation and the variance both increase as r decreases from 1 to 0. In Figure 1, the relationship appears to follow a convex curve.

The Expectation-Variance Efficient Frontier

Figure 1 shows the efficient frontier for the contests we analyzed. Each point represents a contest design. Every point on a line corresponds to a contest with particular parameters.

Figure 1: Expectation and Variance in Contest Designs

Figure 2: Point Z on the Expectation-Variance Frontier

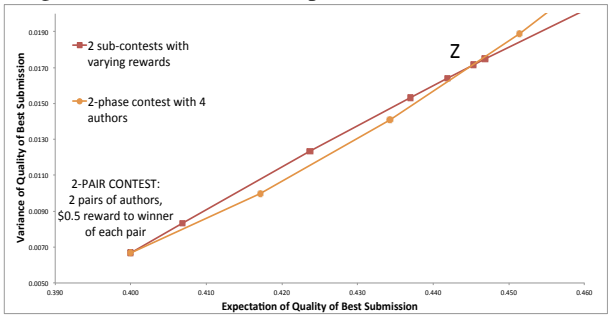

Figure 1 shows that the curves for designs 3 and 4 intersect at a point, which we denote by Z. Figure 2 zooms in on

the area around Z. The Pareto-efficient frontier consists of design 4 to the left of Z and design 3 to the right of Z.

Summary of Theoretical Predictions: In terms of the principal's expected utility alone, a simple contest with 2 authors dominates all other contest designs. When considering *both* the expectation and the variance of the principal's utility, as shown in Figure 2, the picture is more complicated. When the principal's expected utility is important, design 3 to the right and above Z is superior with the pair contest as the best design. However, when a low variance of the principal's utility is desired, design 4 to the left and below Z is superior with the 2-pair contest as the best design. [1]

Experiments on AMT

We implemented the author-critic framework on AMT to compare contest designs and to test our theoretical predictions. To verify that the author-critic framework indeed allows the principal to collect high quality solutions, we compared the pair contest with a baseline design where a *constant* payment is offered for each author or critic task. Moreover, we compared the performances of the pair contest, the quad contest, and the 2-pair contest. Our analysis predicts that the pair contest achieves a higher expectation and a lower variance of the principal's utility than the quad contest, and that the 2-pair contest achieves a lower expectation and a lower variance than the pair and quad contest.

Experimental Design

The task we used was writing a short tourism ad for a city. Our tests were conducted on AMT, so a single author task was to write an ad, and a single critic task was to evaluate an ad. We varied the payment rules for the author tasks to test the theoretical predictions.

Author Task Design Authors were asked to write a short tourism ad for the city of New Orleans, USA or Nice, France, of no more than 600 characters (roughly 100 words). For the constant payment scheme, each author was paid a constant amount for any valid ad. We required a valid ad to be non-empty, original, written in English, relevant to the city specified, and following the length limit. For the contest schemes, each author was paid a base payment for completing the task, and a bonus payment was given to the ad winning the contest. Authors were told that 30 critics would evaluate each ad, and that the bonus would be given to an author if the critics evaluated it to be of the highest quality among all ads. The submitted ads were checked for their originality, and solutions with more than 20% of their content plagiarized from another website were rejected.

Critic Task Design Each critic task includes descriptions of a base payment for competing the task, the bonus payment rule based on the WP method (Witkowski and Parkes 2012) and multiple choice questions to implement the WP method. We briefly described how we collected the ads, then asked each critic to provide their prior expectations regarding quality of ads in our pool, as a numerical value between 0 (lowest quality) and 10 (highest quality). Then we showed the ad to be evaluated, asked each critic whether reading this ad has changed her opinion regarding the quality of ads in our pool, and asked her to provide a new numerical value representing her revised expectation of the quality of ads in our pool. Then the quality of each ad was computed to be the fraction of critics who increased their expectation of the quality of our ads by reading the ad.

Pair Contest versus Constant Payment

We first compared the pair contest with the constant payment scheme. We used the city of New Orleans, USA as our destination. The budget for each chosen ad was fixed to be \$2. For the constant payment scheme, each of 16 authors from AMT were paid \$2 as long as he/she submitted a valid ad. For the pair contest, 32 authors were randomly divided into 16 pairs. Each author was awarded a base payment \$0.4 for submitting a valid ad[2]. Each of the 48 submitted ads was evaluated by 30 critics. The winner of each pair contest was paid a bonus of \$1.2. A basic game theoretic analysis predicts that authors would exert no effort in the constant payment scheme and would exert a positive amount of effort in the pair contest.

Results: Our results show that the pair contest achieved significantly higher expected utility for the principal by Wilcoxon's rank-sum test (MWM-test) ($p < 1\%$) and significantly different variance of the principal's utility by Levene's test ($p < 2\%$) than the constant payment scheme. This suggests that even the simple pair contest scheme can effectively incentivize authors to exert higher effort than a constant payment scheme. However, our results are much less extreme than the theoretical prediction that authors would exert no effort in the constant payment scheme.

Our observations suggest that a sense of "work ethics" strongly influences the behaviors of many workers on AMT (Turkers), similarly to psychological effects described in (Kittur, Chi, and Suh 2008; Paolacci, Chandler, and Ipeirotis 2010). Even for a constant payment, many Turkers submitted ads of considerable quality, and several critics commented that the authors of ads of poor quality did not perform their tasks reasonably and did not deserve to be paid. Consequently, it is reasonable to expect the Turkers' behaviors not to conform to extreme game theoretic predictions.

Pair Contest versus Quad Contest

We now examine the effects of varying the number of authors on the principal's utility in a simple n-author contest. We compared the performance of the pair and quad contests for the city of Nice, France. For the pair contest, 40 authors were randomly divided into 20 pairs. For the quad contest,

[1] For our efficient frontier, we focused on symmetric equilibria of symmetric contests with 2 or 4 authors. There are few designs falling in this class. Contests with *more authors* may be more efficient, especially for non-symmetric contests and non-symmetric equilibria. This is an interesting direction for future research.

[2] We initially aimed to pay the entire \$2 amount only as a bonus to the winner of the contest. However, under this payment scheme it turned out impossible to recruit authors, so we had to allocate some of the \$2 budget for an ad to a base payment.

80 authors were randomly divided into 20 groups of 4 authors each. In both schemes, each author was awarded a base payment of \$0.6 for submitting a valid ad. The winner of each pair/quad contest was paid a bonus of \$0.8. Each of the 120 submitted ads was evaluated by 30 critics. The bonus for each winner is the same, so our theoretical analysis predicts that the principal's expected utility in the quad contest would be lower than that of the pair contest.

Results: Surprisingly, the principal's expected utility of 0.8617 in the quad contest was significantly *higher* than 0.7725 in the pair contest, by MWM-test ($p < 1\%$)[3]. Further, the variance 0.0107 of the principal's utility in the pair contest is higher than the variance 0.0053 in the quad contest, but the difference is not statistically significant by Levene's test ($p > 10\%$). These results contradict our theoretical predictions that the pair contest would outperform the quad contest in terms of the principal's utility. This may due to the relationship between psychological pressure, which depends on the number of competing authors, and performance (Baumeister 1984; Lewis and Linder 1997; Beilock et al. 2004). Further research is required to better characterize this effect and determine the optimal contest structure in practice.

2-Pair Contest versus Pair and Quad Contests

We evaluated the 2-pair contest against the pair and quad contests for the city of Nice, France. We divided 80 authors into 20 groups of 4 people each, and each group participated in a 2-pair contest. Each author was paid \$0.6 for submitting a valid ad. Each group of 4 authors was divided into 2 pairs, and the winner of each pair was paid a bonus of \$0.4. Then, the best ad among the 4 was chosen as the winning ad, though no additional bonus was awarded. Each of the 80 submitted ads was evaluated by 30 critics. Our theoretical analysis predicts that the 2-pair contest would achieve lower expectation and lower variance of the principal's utility than either the pair contest or the quad contest.

Results: The 2-pair contest achieved a expectation of 0.9 and a variance of 0.002 of the principal's utility. By the MWM-test, the expectation of 0.9 in the 2-pair contest is significantly higher than the expectation of 0.7725 in the pair contest ($p < 1\%$). Also, MWM-test shows that the expectation of 0.9 in the 2-pair contest is significantly higher than that of the quad contest ($p < 5\%$). The variance of 0.002 in the 2-pair contest is significantly lower than the variance of 0.0107 in the pair contest ($p < 1\%$, Levene's test). However, the difference in the variances between the 2-pair contest and the quad contest is not significant ($p > 10\%$, Levene's test).

These results suggest that the 2-pair contest is superior to the pair contest, as it achieves higher expectation and lower variance. There is also evidence suggesting that the 2-pair contest is slightly better than the quad contest, though the difference in the variance is not statistically significant. Thus, the 2-pair contest appears to be the best design of all those empirically evaluated.

[3]Note that this value differs from the principal's expected utility in the pair contest in the previous experiment. This may be due to selecting the city of Nice, France, rather than New Orleans.

Conclusions

We theoretically and empirically analyze the author-critic framework based on crowdsourcing contests and peer prediction methods. Our results show that such a scheme is a powerful crowdsourcing tool and outperforms designs offering constant payments. We theoretically characterize a tradeoff between the expectation and the variance the principal's utility for different contest designs. The Pareto efficient frontier shows that the pair contest achieves the highest expectation and the 2-pair contest achieves the lowest variance among all the designs analyzed.

However, our empirical study shows somewhat different results. Most notably, as opposed to the theoretical predictions, our empirical results suggest that increasing the number of contest participants beyond two may improve the principal's utility. Further, using the 2-pair contest achieves the highest expected utility for the principal of all the designs tested, as well as the lowest variance (significantly lower than that of the pair contest and comparable to that of the quad contest). Since both our theoretical and empirical results show that the 2-pair contest achieves a lower variance than the pair/quad contest, and our empirical results show that it achieves the highest expectation, we recommend using the 2-pair contest in practice.

Various factors may account for the differences between the theoretical and empirical results, such as psychological effects and task formulation issues (Kittur, Chi, and Suh 2008; Paolacci, Chandler, and Ipeirotis 2010) or Turkers' desire to reach a target earning (Horton and Chilton 2010), but further work is required to build a more satisfying and complete model. The reasonable quality obtained even in under constant payments indicates that some Turkers have a sense of "work ethics", compelling them to exert effort, even if this does not affect their payment. It also appears that Turkers' effort towards a task may depend on how enjoyable it is. There may be several groups of Turkers: those who exert a lot of effort with little regard to the bonus payment offered, and those who are more "rational" in the game theoretic sense and whose main motivation is monetary. Our author-critic framework is likely to increase the efforts exerted by the latter group and thus improve the quality of the results.

Many questions are also left open for future research. First, what contest structures will maximize the expectation and minimize the variance of the principal's utility in practice? Could models from behavioral game theory account for the differences between our theoretical and empirical results? What is the optimal peer prediction method to use in the author-critic framework? How do mergers and stable collusion schemes, such as those examined in (Huck, Konrad, and Muller 2002; Jost and Van der Velden 2006; Bachrach 2010; Bachrach, Key, and Zadimoghaddam 2010), affect our choice of optimal content design? Finally, are there crowdsourcing contests that improve the efficient expectation-variance frontier for the class of contests we analyzed?

References

Archak, N., and Sundararajan, A. 2009. Optimal design of crowdsourcing contests. *ICIS 2009 Proceedings* 200.

Archak, N. 2010. Money, glory and cheap talk: analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on topcoder. com. In *Proceedings of the 19th international conference on World wide web*, 21–30. ACM.

Bachrach, Y.; Key, P.; and Zadimoghaddam, M. 2010. Collusion in vcg path procurement auctions. *Internet and Network Economics* 38–49.

Bachrach, Y. 2010. Honor among thieves: collusion in multi-unit auctions. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, 617–624. International Foundation for Autonomous Agents and Multiagent Systems.

Baumeister, R. 1984. Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of personality and social psychology* 46(3):610.

Beilock, S.; Kulp, C.; Holt, L.; and Carr, T. 2004. More on the fragility of performance: choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General* 133(4):584.

Boudreau, K.; Lacetera, N.; and Lakhani, K. 2011. Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science* 57(5):843–863.

Chawla, S.; Hartline, J. D.; and Sivan, B. 2011. Optimal crowdsourcing contests. *Workshop on Social Computing and User Generated Content*.

DiPalantino, D., and Vojnovic, M. 2009. Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, EC '09, 119–128. New York, NY, USA: ACM.

Gneezy, U., and Smorodinsky, R. 2006. All-pay auctions–an experimental study. *Journal of Economic Behavior & Organization* 61(2):255–275.

Horton, J. J., and Chilton, L. B. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*, EC '10, 209–218. New York, NY, USA: ACM.

Huck, S.; Konrad, K.; and Muller, W. 2002. Merger and collusion in contests. *Journal of Institutional and Theoretical Economics JITE* 158(4):563–575.

Jost, P., and Van der Velden, C. 2006. Mergers in patent contest models with synergies and spillovers. *Schmalenbach Business Review, Vol. 58, pp. 157-179, April 2006*.

Jurca, R., and Faltings, B. 2005. Enforcing truthful strategies in incentive compatible reputation mechanisms. In Deng, X., and Ye, Y., eds., *WINE*, volume 3828 of *Lecture Notes in Computer Science*, 268–277. Springer.

Jurca, R., and Faltings, B. 2007. Robust incentive-compatible feedback payments. In *Proceedings of the 2006 AAMAS workshop and TADA/AMEC 2006 conference on Agent-mediated electronic commerce: automated negotiation and strategy design for electronic markets*, TADA/AMEC'06, 204–218. Berlin, Heidelberg: Springer-Verlag.

Jurca, R., and Faltings, B. 2009. Mechanisms for making crowds truthful. *J. Artif. Intell. Res. (JAIR)* 34:209–253.

Kittur, A.; Chi, E.; and Suh, B. 2008. Crowdsourcing user studies with mechanical turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 453–456. ACM.

Lewis, B., and Linder, D. 1997. Thinking about choking? attentional processes and paradoxical performance. *Personality and Social Psychology Bulletin* 23(9):937–944.

Liu, T.; Yang, J.; Adamic, L.; and Chen, Y. 2011. Crowdsourcing with all-pay auctions: a field experiment on taskcn.

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting informative feedback: The peer-prediction method. *Manage. Sci.* 51:1359–1373.

Moldovanu, B., and Sela, A. 2001. The optimal allocation of prizes in contests. *American Economic Review* 91(3):542–558.

Paolacci, G.; Chandler, J.; and Ipeirotis, P. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making* 5(5):411–419.

Prelec, D. 2004. A bayesian truth serum for subjective data. *Science* 306.

Siegel, R. 2009. All-pay contests. *Econometrica* 77(1):71–92.

Witkowski, J., and Parkes, D. 2012. Peer Prediction Without a Common Prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)*.