WiseMarket: A New Paradigm for Managing Wisdom of Online Social Users

Caleb Chen Cao [†] Yongxin Tong [†] Lei Chen [†] H. V. Jagadish [‡]
[†]The Hong Kong University of Science and Technology, Hong Kong SAR, PR China
[‡]University of Michigan, Ann Arbor, MI, USA
{caochen,yxtong,leichen}@cse.ust.hk, [‡]jag@umich.edu

ABSTRACT

The benefits of crowdsourcing are well-recognized today for an increasingly broad range of problems. Meanwhile, the rapid development of social media makes it possible to seek the wisdom of a crowd of targeted users. However, it is not trivial to implement the crowdsourcing platform on social media, specifically to make social media users as workers, we need to address the following two challenges: 1) how to motivate users to participate in tasks, and 2) how to choose users for a task. In this paper, we present *Wise Market* as an effective framework for crowdsourcing on social media that motivates users to participate in a task with care and correctly aggregates their opinions on pairwise choice problems. The *Wise Market* consists of a set of *investors* each with an associated individual confidence in his/her prediction, and after the investment, only the ones whose choices are the same as the whole market are granted rewards. Therefore, a social media user has to give his/her ''best'' answer in order to get rewards, as a consequence, careless answers from sloppy users are discouraged.

Under the *Wise Market* framework, we define an optimization problem to minimize expected cost of paying out rewards while guaranteeing a minimum confidence level, called the *Effective Market Problem (EMP)*. We propose exact algorithms for calculating the market confidence and the expected cost with $\mathcal{O}(n \log^2 n)$ time cost in a *Wise Market* with n investors. To deal with the enormous number of users on social media, we design a Central Limit Theorem-based approximation algorithm to compute the market confidence with $\mathcal{O}(n)$ time cost, as well as a bounded approximation algorithm to calculate the expected cost with $\mathcal{O}(n)$ time cost. Finally, we have conducted extensive experiments to validate effectiveness of the proposed algorithms on real and synthetic data.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications---*Data mining*; H.1.2 [**MODELS AND PRINCIPLES**]: User/Machine Systems---*Human information processing*

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

Keywords

Crowdsourcing, Human Computation, Social Media, Market

1. INTRODUCTION

The wisdom of crowds has long been known, but only with the development of online services has it becomes possible to enroll and manage the crowd's work effectively and handle payments efficiently. On traditional crowdsourcing platforms like Amazon Mechanical Turk (AMT), task holders post their tasks onto a public task pool and wait for workers to choose them. Once the tasks are completed, workers are paid via the crowdsourcing platform. These centralized crowdsourcing platforms contribute greatly to facilitate the power of crowds. However, there are certain natural drawbacks of this mechanism: 1) Payments are only weakly associated with the quality of the work from crowd workers, so some workers try to maximize income by giving answers fast, even to the extent of giving random answers; 2) Task holders cannot actively choose the workers. Rather they have to accept anyone who shows up and meets the published criteria, thereby having very limited control on the quality of workers recruited, and hence on the quality of work produced; And 3) The crowd tends to be from a restricted demographic class [10], which may produce biased results when demographics matter.

Users of social media services can be a huge reservoir of workers for crowdsourcing: 1) Users of social media range over a much broader demographics, and of course, a larger population; 2)Information in the users' profile can be used to infer some of their interests and abilities; 3) Task holders can actively enroll targeted workers using built-in functions like '@' (in Tiwtter) or private messages (in Facebook). Therefore, there is an emerging notion of "managed crowdsourcing" [5, 6], where workers are actively chosen. In this paper, we present *WiseMarket* as an effective framework to support such managed crowdsourcing, for tasks are two-option decision making problems. To this end, we tackle two essential challenges: how to motivate users to participate in tasks, and how to choose users for tasks.

Markets are known to be an effective institution for aggregating beliefs of online users and yielding reliable answers. For example, in Racetrack Betting [8] investors are only allowed to choose from two options, and the promised rewards are only given to the investors whose investment meets the market opinion (the majority of all investors). We use this idea as the basis for our framework. Online users are considered investors in this market, while the social media is treated as a collective knowledge base. Thus, a wise market includes a set of investors, each associated with a probability (*individual confidence*) that he or she will give the correct answer. The system maintains a pool of candidate users while evaluating the users' confidence simultaneously. When a particu-

Table 1: 8 cases in WM_ACD

Case	Correct	Wrong	Prob.	Cost
1	{A,C,D}	$\emptyset$	0.448	3
2	{A,C}	{D}	0.112	2
3	{A,D}	{C}	0.192	2
4	{C,D}	{A}	0.112	2
5	{A}	{C,D}	0.048	2
6	{C}	{A,D}	0.028	2
7	{D}	{A,C}	0.048	2
8	$\emptyset$	{A,C,D}	0.012	3

lar task comes, the system recommends an optimal subset of users and releases the task to them with promise of a reward. After their choices, the answer preferred by a majority of the users will be output as the market opinion and the ones who make the same choice as the market are granted a reward. Note that the output is the majority of the market, not necessarily the same as the ground truth. The probability that the market opinion meets the ground truth is considered as the *Market Confidence*. The expected total reward payout is the *Market Cost* borne by the task-owner. Then the problem of building an ''effective'' market is to find a set of investors from all possible users, whose *Market Cost* is minimum subject to its *Market Confidence* satisfying a given threshold. We show the problem with the following example.

EXAMPLE 1. *Consider a candidate set of investors (social media users)* $A(0.8), B(0.6), C(0.7), D(0.8), E(0.5)$. *Each candidate investor has an individual confidence as in brackets, and the task-owner has a market confidence threshold* $\theta = 0.85$. *Our problem is to find a set of investors who have the minimum market cost subject to the market confidence threshold.*

First, we try the market, $WM_1 = \{A(0.8), C(0.7), D(0.8)\}$, *following the majority rule,* WM_1 *works correctly only when at least two investors make the correct choice. Specifically, there are in total four out of eight cases that the market provides the correct answers as shown in Table 1. In the first 4 cases, the number of correct investors are in a majority(i.e. no less than 2). We obtain the probability that the entire market performs correctly(*Market Confidence*) by summing up the probabilities of the first 4 cases, i.e.* $MC(WM_1) = 0.448 + 0.112 + 0.192 + 0.1912 = 0.864$.

Turning to the cost, we need to pay three units for case 1, and two units each for cases 2,3,4. Note that even when the market does not give the correct answer as in the last 4 cases, payments still need to be made to the majority of the investors. We denote the investors in majority as a Winning Set*, whose size determines the payment. The expected value of cost* $E[Cost] = 3 \cdot \sum$(*Prob. of Case 1 and 8*) $+$ $2 \cdot \sum$(*Prob. of Case 2 to 7*) $= 2.46$.

If we instead selected $WM_2 = \{B(0.6), C(0.7), D(0.8)\}$, *it turns out that* WM_2 *has a market confidence of 0.788 and an expected cost of* 2.36. WM_2 *is more ''cheaper'' than* WM_1. *However, since the given threshold is* $\theta = 0.85$, WM_2 *does not satisfy the confidence threshold. In such a case,* WM_1 *is the answer.*

In this paper, to address the problem of building an effective wise market on social media, we have made the following contributions:

- We propose *Wise Market*, a new framework to manage the wisdom of online social users.
- We define the *Effective Market Problem* as the central design challenge for effective utilization of wise markets.
- We design efficient exact algorithms to calculate the market confidence and the expected cost, as well as a Central Limit Theorem-based approximation algorithm for the market confidence and an approximation algorithm for the expected cost with an approximation ratio of $(3 - 2\theta)$.
- We provide an efficient algorithm to solve the *Effective Market Problem*, which seamlessly incorporates the proposed algorithms and two effective pruning methods.

The rest of the paper is organized as follows. A formal definition of the problem is presented in Section 2. In Section 3, we study the calculation of Market Confidence and present both exact and approximate algorithms and in Section 4, we study the structure of calculating Market Cost and propose both exact and approximation algorithms. Then, in Section 5, we build upon the work in the preceding two sections to present the overall Effective Market Algorithm(EMA). In Section 6, we empirically study the intrinsic characteristics of the problem and the performance of the proposed algorithms. Finally, in Section 7, we summarize the most recent related work, and conclude in Section 8.

2. PROBLEM DEFINITION

2.1 Investor

In a *Wise Market*, each $investor$ works on tasks, each of which is to make a choice between two given options. The task may have a (latent) ground truth, but this is not known to us. (Some tasks may have a ground truth that is difficult to characterize -- e.g. which of two logos is more "eye-catching".) When the option selected is the same as the aggregate market choice (not necessarily the latent ground truth), the investor receives a reward of one unit. Investors make decisions based on their experience and general knowledge, and there is a chance that an investor fails to identify the true value of the target. So for an investor ι_i, we define the individual confidence c_i to describe how likely an investor is to choose correctly, and we use another variable $v_i \in \{0, 1\}$ to denote its actual vote between the two choices.

DEFINITION 1 (INVESTOR CONFIDENCE). *For each investor* ι_i, *the* Investor Confidence c_i *is the probability that* ι_i *chooses the same option as the ground truth.*

$$\begin{aligned} c_i &= \Pr\{\iota_i \textit{chooses correctly}\} \\ &= \Pr\{G = 0\} \cdot \Pr\{v_i = 0 | G = 0\} \\ &\quad + \Pr\{G = 1\} \cdot \Pr\{v_i = 1 | G = 1\} \\ &= \Pr\{v_i = G | G\} \end{aligned}$$

where G is the ground truth, the confidence c_i is a probability of making the correct choice, and there are two standard ways to evaluate it. First, confidence could be estimated from the frequency of correct judgments, which could be implemented by analyzing previous records or inserting tasks with known answers [12]. Second, since the probability reflects the subjective degree of the belief, the confidence c_i can also be estimated by associating the relative authoritativeness of users [6].

2.2 Wise Market

A *Wise Market* is a set of investors along with rules to aggregate the market information and issue rewards to successful investors. Assume $I = \{\iota_1, \iota_2, \ldots, \iota_N\}$ is the set of all investors, we define a *Wise Market* as follows:

DEFINITION 2 (*Wise Market*). *A* Wise Market *is a set of investors* $WM_n = \{\iota_1, \iota_2, \ldots, \iota_n\} \subseteq I$ *with size* n, *where each* ι_i *is associated with an individual confidence* c_i *and actual vote* v_i.

The most popular aggregation rule is *Majority Voting*, which presents the opinions of the majority of investors as output. We denote such output as the Market Opinion and present the formal definition as follows:

DEFINITION 3 (MARKET OPINION). *Given a* Wise Market WM, *the Market Opinion* $OP(WM_n)$ *is the aggregated result*

according to the following equation:

$$OP(WM_n) = \begin{cases} 1 & if \sum v_i \geq \lceil \frac{n}{2} \rceil \\ 0 & if \sum v_i \leq \lfloor \frac{n}{2} \rfloor \end{cases}$$

To avoid the cumbersome special case where the vote is tied, in this paper we study the case where the size of a *Wise Market* is **ODD**.

2.2.1 Market Confidence

As an institution for aggregating the distributed knowledge, the quality of a *Wise Market* is measured according to the probability that the market successfully identifies the true value of a task. For better illustration, we first define the concept of *Truth Set* $C = \{\iota_i | \iota_i \in WM_n \text{ s.t. } v_i = G\}$, which includes the investors whose answers are identical to the ground truth G.

DEFINITION 4 (MARKET CONFIDENCE). *The Market Confidence MC is defined as the probability that the Market Opinion is the same as the ground truth G:*

$$\begin{aligned} MC(WM_n) &= \Pr(OP(WM_n) = G|G) \\ &= \Pr(\big|C\big| \geq \lceil \frac{n}{2} \rceil) = \Pr(\big|C\big| \geq \frac{n+1}{2}) \\ &= \sum_{k=\lceil \frac{n}{2} \rceil}^{n} \sum_{A \in F_k} \prod_{i \in A} c_i \prod_{j \in A^c} (1 - c_j) \end{aligned}$$

where $F_k = \{A \big| |A| = k, A \subseteq WM_n\}$ is all the subsets of WM_n with size k and A^c is the complementary set of A.

2.2.2 Market Cost

Another significant concept for a *Wise Market* is the *Market Cost*. We assume that unit rewards are granted to the winning investors. In such a setting, the market cost is numerically equal to the number of winning investors in a market, in other words, the size of the majority in a market.

We define the concept of *Winning Set* as the set of investors who have the same opinion as the market opinion. We then formally define the Market Cost as follows:

DEFINITION 5 (MARKET COST). *Given a* Wise Market WM_n*, the Market Cost, $Cost(WM_n)$, is the size of the* Winning Set*:*

$$Cost(WM_n) = \big|W\big| = \Big|\{\iota_i | \iota_i \in WM_n \text{ s.t. } v_i = OP(WM_n)\}\Big|$$

Note that, only when $OP(WM_n) = G$, the two sets W and C are equal, otherwise $|W| + |C| = n$.

The *Market Cost* varies according to different constitution of *Wise Market*, and to measure the effectiveness of a *Wise Market*, we present the concept of Expected Market Cost, which follows the definition of expected value:

$$\begin{aligned} & E[Cost(WM_n)] \\ &= \sum_{k=\lceil \frac{n}{2} \rceil}^{n} k \cdot \Pr(|W| = k) \\ &= \sum_{k=\lceil \frac{n}{2} \rceil}^{n} k \cdot [\sum_{A \in F_k} \prod_{i \in A} c_i \prod_{j \in A^c} (1 - c_j) \\ &\qquad + \sum_{A \in F_k} \prod_{i \in A} (1 - c_i) \prod_{j \in A^c} c_j] \end{aligned}$$

2.3 Effective Market Problem

Formally, we define the Effective Market Problem as follows:

DEFINITION 6 (EFFECTIVE MARKET PROBLEM). *Given a set of investors $I = \{\iota_1, \ldots, \iota_N\}$ with size N, a Market Confidence threshold θ, the Effective Market Problem(*EMP*) is to find a subset of all investors $WM_n \subseteq I$, so that:*

$$\begin{aligned} \text{minimize} \quad & E[Cost(WM_n)] \\ \text{subject to} \quad & MC(WM_n) \geq \theta \end{aligned}$$

3. CALCULATION OF MARKET CONF

There are three main subproblems to be studied to build up an effective *Wise Market* from a given set of investors, the first one is to efficiently calculate the Market Confidence, the second one is to calculate the Market Cost, and the third one is to efficiently pinpoint the most effective subset of all investors as a market. We study each of these sub-problems in turn in the following three subsections.

3.1 A Divide-and-Conquer-based Exact Method

Based on Definition 4, the Market Confidence aims to compute the probability that the size of C is no less than $\lceil \frac{n}{2} \rceil$. Actually, the size of C ($\big|C\big|$) is considered to be a discrete random variable. Therefore, the probability mass function of $\big|C\big|$ is shown as follows,

$$\Pr(\big|C\big| = k) = \sum_{A \in F_k} \prod_{i \in A} c_i \prod_{j \in A^c} (1 - c_j)$$

For the corresponding probabilities of different values of $\big|C\big|$, we store them in a vector.

In order to compute the Market Confidence efficiently, we propose a divide-and-conquer-based exact algorithm (Algorithm DC in Figure 1). The algorithm first divides the set of investors WM_n into two groups, WM_{n1} and WM_{n2}, as long as WM_n does not have a single member in lines 1-2. Then, the algorithm recursively computes the probability mass function of $\big|C\big|$ in the two partitioned groups in lines 3-4. As a consequence, we get two vectors of size $\frac{n}{2}$ (when n is even), to store the probability distribution of $\big|C\big|$ in WM_{n1} and WM_{n2}, respectively. In fact, the probability of each possible value of global $\big|C\big|$ can be considered as the convolution of corresponding probabilities in two partitioned groups. According to the concept of convolution, the corresponding probabilities of each value of $\big|C\big|$ in n investors can be represented as:

$$Vec_{|C|}[k] = \sum_{i=0}^{k} Vec_{lower}[i] \times Vec_{upper}[k - i]$$

where Vec_{lower} and Vec_{upper} are used to store the probability distribution of $\big|C\big|$ in WM_{n1} and WM_{n2}, respectively. Moreover, in order to speedup the recursive computation of convolution, we employ the *Fast Fourier Transform (FFT)* and the *Inverse Fast Fourier Transform(iFFT)* in line 5. In addition, the algorithm exits in lines 8-10. Thus, based on Algorithm DC, we can obtain the probability distribution of $\big|C\big|$ and compute $MC(WM_n)$ of the given set of investors WM_n easily. Algorithm MCA(Figure 2) shows the details of computing market confidence of WM_n.

Back to the running example, we illustrate Algorithm DC in Figure 3: We first split the set of $\{A, C, D\}$ into two smaller sets $\{A\}$ and $\{C, D\}$, and recursively split $\{C, D\}$ into $\{C\}$ and $\{D\}$. Then we merge the smaller sets using convolution, and *FFT* and *iFFT* are adopted to speedup the process.

Computational Complexity Analysis: In line 5 of Algorithm DC, the computation cost of the FFT-based convolution is $\mathcal{O}(n \log n)$, where n is the size of WM_n. Thus, the whole complexity of Algorithm DC is $\mathcal{O}(n \log^2 n)$. Consequently, the exact market confidence can be computed by Algorithm MCA with computational complexity of $\mathcal{O}(n \log^2 n)$.

Algorithm DC {
Input: A set of investors WM_n
Output: A vector of probability distribution of $|C|$, $Vec_{|C|}$
(1) if ($n \neq 1$)
(2) split n investors into two groups $WM_{\lceil \frac{n}{2} \rceil}$ and $WM_{\lfloor \frac{n}{2} \rfloor}$;
(3) $Vec_{upper} \leftarrow DC(WM_{\lceil \frac{n}{2} \rceil})$;
(4) $Vec_{lower} \leftarrow DC(WM_{\lfloor \frac{n}{2} \rfloor})$;
(5) $Vec_{|C|} \leftarrow iFFT(FFT(Vec_{upper}) * FFT(Vec_{lower}))$;
(6) return $V_{|C|}$;
(7) else
(8) $Vec_{|C|}[0] \leftarrow 1 - c_1$;
(9) $Vec_{|C|}[1] \leftarrow c_1$;
(10) return $Vec_{|C|}$;
}

Figure 1: Divide-and-Conquer-based Algorithm (DC)

Algorithm MCA {
Input: A set of investors WM_n
Output: The market confidence $MC(WM_n)$
(1) $Vec_{|C|} \leftarrow DC(WM_n)$);
(2) for $i \leftarrow \lfloor \frac{n}{2} \rfloor + 1$ to n
(3) $MC(WM_n) \leftarrow MC(WM_n) + Vec_{|C|}[i]$;
(4) return MC;
}

Figure 2: Market Confidence Algorithm (MCA)

3.2 A Chernoff-Inequality-based Bounding

According to Definition 6, we only need to test whether the market confidence of WM_n is larger than the given threshold and need not to get the exact value of $MC(WM_n)$. Thus, we are interested only in an efficient checking method. In this part, we include a tight upper bound of $MC(WM_n)$ to help speed up the checking.

LEMMA 1. *(Chernoff-Inequality-based Bounding) Given a* Wise Market WM_n*, a market confidence threshold* θ*,* $MC(WM_n)$ *satisfies the following upper bound,*

$$MC(WM_n) < \begin{cases} 2^{-\delta\epsilon} & \delta > 2e-1 \\ e^{-\frac{\delta^2\epsilon}{4}} & 0 < \delta < 2e-1 \end{cases}$$

where $\epsilon = \sum_{\iota_i \in WM_n} c_i$ *and* $\delta = (\lceil \frac{n}{2} \rceil - \epsilon)/\epsilon$.

Note that due to the limit of the space, proofs of all the lemmas in this paper can be found in the technical report [7].

Based on Lemma 1, we only spend $\mathcal{O}(n)$ time to check whether the upper bound satisfies the confidence threshold.

3.3 A Central-Limit-Theorem-based Approximation Algorithm

As we discussed above, the Chernoff-Inequality-based bounding technique of $MC(WM_n)$ can improve the efficiency of the exact solution by employing the upper bound to check whether the set can be pruned. Then, we perform Algorithm MCA if the set fails to be filtered. However, in a real social network, which features a large number of candidate users, the computational cost is still high, especially when the set of investors fail to satisfy the market confidence threshold and cannot be filtered by the above pruning method. In this subsection, we propose an efficient Central-Limit-Theorem-based approximation algorithm which not only has linear computational complexity but also returns the result with a high confidence. The basic idea is to use the probability density function of a Standard Normal distribution to replace the probability mass function of $|C|$ according to the Central Limit Theorem. Thus, $MC(WM_n)$ can be approximately computed by the probability density function of a Standard Normal distribution. In the following, we firstly prove that the probability distribution of $|C|$ converges in probability to a Standard Normal distribution, and then propose our approximation algorithm.

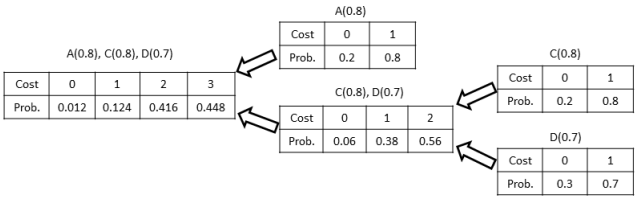


Figure 3: DC Algorithm on Running Example

THEOREM 1. *(Central Limit Theorem of the probability distribution of* $|C|$*) Given a* Wise Market WM_n*,* $MC(WM_n)$ *converges in probability to a Standard Normal distribution, namely*

$$MC(WM_n) \sim \Phi(\frac{\lceil \frac{n}{2} \rceil + 0.5 - \epsilon}{\sigma})$$

where $\epsilon = \sum_{\iota_i \in WM_n} c_i$ *and* $\sigma^2 = \sum_{\iota_i \in WM_n} c_i(1-c_i)$.

PROOF. For each investor associated with individual confidence c_i, we denote ζ_i as the random event that ι_i *invests correctly*, which follows the Binomial distribution. Thus, the expectation and variance of ζ_i are $E(\zeta_i) = c_i$ and $Var(\zeta_i) = c_i(1-c_i)$, respectively. For a given *Wise Market* WM_n, $|C| = \sum_{i=1}^{n} \zeta_i$, so the expectation and variance of $|C|$ are

$$E(|C|) = \sum_{i=1}^{n} c_i \; and \; Var(|C|) = \sum_{i=1}^{n} c_i(1-c_i)$$

Because different ι_i may have different c_i, $\zeta_1 ... \zeta_n$ are independent and may not follow a same distribution. Based on the Lyapunov's Central Limit Theorem, the probability distribution of $|C|$ converges in probability to a Standard Normal distribution *iff* the following two conditions hold:

1. $E[|\zeta_i|]^{2+\delta}$ *is finite where* $\delta > 0 (i = 1, ..., n)$
2. $\lim_{n \to \infty} \frac{1}{Var(|C|)^{2+\delta}} E[|\zeta_i - E(\zeta_i)|]^{2+\delta} = 0$ *if there is* $\delta > 0$

For the first condition, it is obviously correct because $E(\zeta_i)$ and $Var(\zeta_i)$ are finite. And the second condition is also satisfied based on the existence of a suitable δ. A detailed proof can be found in our technical report [7]. □

Based on Theorem 1, we design an approximate algorithm(see Algorithm CLT-MC in Figure 4) to compute $MC(WM_n)$ efficiently. In Algorithm CLT-MC, the expectation ϵ and the variance σ of $|C|$ are accumulated in lines 2-4, respectively. Then, in line 5, the $MC(WM_n)$ can be obtained by a hash function which maps different expectations and variances to the corresponding probabilities under a Standard Normal distribution.

Back to the running example, while $\epsilon = E(|C|) = 0.8 + 0.8 + 0.7 = 2.3$ and $\delta = \sqrt{Var(|C|)} = \sqrt{0.16 + 0.16 + 0.21} = 0.728$, according to Theorem 1, $\phi(\frac{2+0.5-2.3}{0.728}) = 0.61$. Note that as the data size increases, the approximated value gets closer to the real value. Please refer to Figure 8(g) and Figure 8(i).

Computational Complexity Analysis:
In Algorithm CLT-MC, the computational complexity is only $\mathcal{O}(n)$ where n is the size of *Wise Market* WM_n because we only accumulate the expectation and the variance of $|C|$. Moreover, since we implement the probability calculation of the Standard Normal distribution via a hashing function, line 5 only requires $\mathcal{O}(1)time$.

4. CALCULATION OF MARKET COST

Besides testing the market confidence requirement, calculating the market cost is another important task of the *Effective Market*

Algorithm CLT-MC {
Input: A set of investors WM_n
Output: The market confidence $MC(WM_n)$
(1) $MC(WM_n) \leftarrow 0$;
(2) for $i \leftarrow 1$ to n
(3) $\epsilon \leftarrow \epsilon + c_i$;
(4) $\sigma^2 \leftarrow \sigma^2 + (c_i(1-c_i))$;
(5) $MC(WM_n) \leftarrow \Phi(\frac{\lceil \frac{n}{2} \rceil + 0.5 - \epsilon}{\sigma})$;
(6) return $MC(WM_n)$;
}

Figure 4: CLT-based Approximation Market Confidence Algorithm (CLT-MC)

Problem. In this section, we design an exact and an approximation algorithms to compute the market cost, respectively.

4.1 An Exact Algorithm

In this subsection, we focus on how to compute the market cost exactly. Based on Definition 5, we can find that it is easy to compute after the probability distribution of $\left|C\right|$ is known. Thus, the basic idea of the exact algorithm is to compute the expected market cost directly from the probability distribution of $\left|C\right|$. More details are shown in Algorithm EEC(Figure 5).

In Algorithm EEC, it first calls Algorithm DC to obtain the probability distribution of $\left|C\right|$ in line 1. Then, according to Definition 5, the expected market cost is computed in lines 2-6 and the final market cost is returned in line 7.

Computational Complexity Analysis: Because Algorithm EEC employs Algorithm DC in line 1, this step spends $\mathcal{O}(n \log^2 n)$ computational cost where n is the size of the *Wise Market*. In lines 2-6, the algorithm spends $\mathcal{O}(n)$ time cost. Thus, the total computational complexity of Algorithm EEC is $\mathcal{O}(n \log^2 n) + \mathcal{O}(n) = \mathcal{O}(n \log^2 n)$.

4.2 An Approximation Algorithm

In order to compute the market cost efficiently, we also introduce an approximation algorithm which has linear computational complexity and a constant approximation ratio. Before we discuss the approximation algorithm, we introduce a pair of lower and upper bounds of the expected market cost via the following two lemmas.

LEMMA 2. *(Lower Bound of Expected Market Cost) Given a* Wise Market WM_n*, the following inequality holds:*

$$\sum_{i=1}^{n} c_i < E[Cost(WM_n)]$$

where each c_i is the invest confidence of each ι_i.

In addition to the lower bound of the expected market cost in Lemma 2, we also find the upper bound of the expected market cost in the following lemma.

LEMMA 3. *(Upper Bound of Expected Market Cost) Given a* Wise Market WM_n*, and a market confidence threshold θ, the following inequality holds:*

$$E[Cost(WM_n)] < \sum_{i=1}^{n} c_i + n(1-\theta)$$

Based on the aforementioned upper bound of the expected market cost, we can use the upper bound instead of the expected market cost as the approximation result and guarantee that the approximation result is at most $(3-2\theta)$ times larger than the exact result in the following theorem.

THEOREM 2. *(Approximation Ratio of Approximate Expected Market Cost) Given a* Wise Market WM_n*, a market confidence threshold θ, the upper bound in Lemma 3 is at most $3-2\theta$ times larger than the exact expected market cost.*

Algorithm EEC {
Input: A set of investors WM_n
Output: The expected Market Cost $E[Cost(WM_n)]$
(1) $Vec_{|C|} \leftarrow DC(WM_n)$;
(2) for $i \leftarrow 1$ to n
(3) if $i \geq \frac{n}{2}$
(4) $E[Cost(WM_n)] \leftarrow E[Cost(WM_n)] + i \times Vec_{|C|}[i]$;
(5) else
(6) $E[Cost(WM_n)] \leftarrow E[Cost(WM_n)] + (n-i)Vec_{|C|}[i]$;
(7) return $E[Cost(WM_n)]$;
}

Figure 5: Exact Expected Market Cost Algorithm (EEC)

PROOF *(Sketch).* According to the probability distribution of $\left|C\right|$, we can do the following transformation.

$$n = n \cdot 1 = n \cdot \sum_{k=\lceil \frac{n}{2} \rceil}^{n} [\sum_{A \in F_k} \prod_{i \in A} c_i \prod_{j \in A^c} (1-c_j) + \sum_{A \in F_k} \prod_{i \in A} (1-c_i) \prod_{i \in A^c} c_j]$$

Then, we know that

$$2E[Cost(WM_n)] = \sum_{k=\lceil \frac{n}{2} \rceil}^{n} 2k \cdot [\sum_{A \in F_k} \prod_{i \in A} c_i \prod_{j \in A^c} (1-c_j) + \sum_{A \in F_k} \prod_{i \in A} (1-c_i) \prod_{i \in A^c} c_j] > n$$

According to the upper bound of expected market cost in Lemma 3,

$$\begin{aligned} E[Cost(WM_n)] &< \sum_{i=1}^{n} c_i + n(1-\theta) \\ &< \sum_{i=1}^{n} c_i + 2E[Cost(WM_n)] \cdot (1-\theta) \\ &< E[Cost(WM_n)] + 2(1-\theta) \cdot E[Cost(WM_n)] \ (Lemma\ 2) \\ &= (3-2\theta) \cdot E[Cost(WM_n)] \end{aligned}$$

Therefore, the upper bound of the expected market cost is at most $(3-2\theta)$ times of the exact result. □

Based on Theorem 2, we can design an approximation algorithm which has $(3-2\theta)-$approximation ratio in Algorithm AEC(Figure 6).

In Algorithm AEC, we firstly compute ϵ, the expectation of $\left|C\right|$ in lines 1-2. Then, we return an approximation result in line 3.

Back to the running example, according to the results in Section 3(Figure 3), we can obtain the exact cost $E[Cost(WM_1)] = 0.012 \cdot 3 + 0.124 \cdot 2 + 0.416 \cdot 2 + 0.448 \cdot 3 = 2.46$. While using Algorithm 6, we could obtain an approximate value of $E[Cost(WM_1)] = 2.3 + 3 \cdot (1 - 0.85) = 2.75$, where the threshold $\theta = 0.85$ and the algorithm achieves an approximation ratio equal to 1.1.

Computational Complexity Analysis:

Because Algorithm AEC only needs to compute the expectation of $\left|C\right|$, the total computational complexity of Algorithm AEC is $\mathcal{O}(n)$.

5. BUILDING AN EFFECTIVE MARKET

In this section, we discuss the strategy to solve the effective market problem by integrating both algorithms of computing the market confidence and the expected market cost. According to the definition of the expected market cost, there is not any monotonicity or anti-monotonicity between the subset/superset relationship. So we have to enumerate all possible subsets of the N investors and then check their market confidences and the expected market costs via our solutions discussed above. Therefore, an effective and efficient heuristic algorithm for the effective market problem is shown by the following filtering-and-verification framework.

Algorithm AEC {
Input: A set of investors WM_n
Output: The expected Market Cost $E[Cost(WM_n)]$
(1) for $i \leftarrow 1$ to n
(2) $\epsilon \leftarrow \epsilon + c_i$;
(3) return $E[Cost(WM_n)] \leftarrow \epsilon + n(1-\theta)$;
}

Figure 6: Approximation Expected Market Cost Algorithm (AEC)

First Step: We first check whether there exists at least one investor whose market confidence is no less than the market confidence threshold θ. If so, the effective market only includes the investor who has the smallest but threshold-satisfying confidence among all the given investors, and then the algorithm stops. Actually, it is equal for any eligible investor to be selected when the size of the effective market is one. However, for the case that the size of the effective market is more than one, an effective heuristic rule is that invertors with the smallest confidence are tried firstly because the unqualified markets can be safely pruned in $\mathcal{O}(n)$ time complexity (Lemma 1 and Algorithm DC). On the contrary, we have to spend $\mathcal{O}(nlog^2n)$ time complexity to test the unqualified markets. Therefore, we still select the investor with the smallest confidence when the size of market is one to keep the consistent strategy style.

Second Step: If there is no single investor as the effective market, we will enumerate all possible subsets. The main heuristic rule is to search from the smallest market to the larger ones, and test from the combination of investors with smallest confidence to that with higher confidence. Thus, if the Chernoff inequality-based upper bound of a subset is smaller than θ, this subset can be filtered.(**Filtering Phase**)

Third Step: For the subsets of all investors that cannot be pruned, we compute their expected market costs and find the minimum one. (**Verification Phase**)

Note that in the first step, we select the investors with the smallest confidence for quickly finding the eligible crowd with the minimum number of investors. Based on the framework above, in the worst case, this solution is $\mathcal{O}(2^N)$ where N is the size of all the given investors. For a large amount of data, it is expensive to use this exact method to find the effective market. Therefore, according to the approximation algorithm of the expected market cost in Algorithm AEC, we propose an efficient approximation algorithm to discover the effective market in Algorithm EMA(Figure 7). The basic idea is to search for an effective market in the increasing order of size. For the subsets of same size, we can rank them by their approximation expected Market Cost (Algorithm AEC) in an ascending order. Once we find a subset whose market confidence is greater than θ, we stop and return this subset as the final effective market. More details can be found in Algorithm EMA. Before discussion of Algorithm EMA, we propose a pruning method called *Early Termination* to speed up the algorithm.

LEMMA 4. *(Early Termination) Given a set of investors $I = \{\iota_1, \iota_2, \ldots, \iota_N\}$, a subset $M \subseteq I$ which has the current minimum market cost(CMC) so far, a market confidence threshold θ, the algorithm of* EMP *can stop if and only if $CMC \leq MC(M'_{min})$ where M'_{min} contains $|M|+2$ investors with the smallest individual confidence c_i.*

In Algorithm EMA, we firstly check whether the largest probability among N investors is larger than the confidence threshold θ. If so, we can directly return the investor who has the minimum probability among investors whose probabilities are not less than θ in line 1. Otherwise, we continue to check other subsets of the N investors in lines 3-12. We first perform the *Early Termination* pruning of Lemma 4 in lines 5-6. Then, for the set of all subsets whose size is i, S_i , we use the function $RankMerge(S_i)$ to generate all (i+2)-size subsets by merging any two i-size subsets and to

Algorithm EMA {
Input: A set of candidate investors $I = \{\iota_1, \ldots, \iota_N\}$, a market confidence threshold θ
Output: the effective market EM
(1) if $(Max(c_i) \geq \theta)$
(2) return $EM \leftarrow c \in WM_n$ *s.t. minimized{E[Cost(c)]} and* $MC(c) \geq \theta$;
(3) else
(4) for $i \leftarrow 1$ to $n-1$
(5) If $EM \leq [AEC(min(s_{i+2}), \theta)]$
(6) return EM;
(7) $M_{i+2} \leftarrow RankMerge(M_i)$;
(8) for $(s_j \in M_{i+2})$
(9) if $MC(s_j) \geq \theta$ && $AEC(s_j, \theta) \leq EM$
(10) $EM \leftarrow AEC(s_j, 0)$;
(11) else
(12) continue;
}

Figure 7: Effective Market Algorithm (EMA)

rank them according to the combinatorial order of their ranked confidences in line 7, e.g. $< c_1, c_2, c_5 >$ ranks before $< c_1, c_3, c_4 >$. If there exists at least one subset whose market confidence is no less than θ, we find the subsets which have the smallest expected market cost among all (i+2)-size subsets whose market confidences are no less than θ (lines 8-10). In addition, in lines 11-12, we directly terminate the process to check any (i+2)-size subset if all i-size subsets fail to satisfy the confidence threshold θ .

6. EVALUATION

In this section, we present a series of empirical studies of the performance of our proposed algorithms. All experiments are conducted on a PC with 2 Intel(R) Core(TM) 2.13GHz CPU and 2GB memory, running on Microsoft 64-bit Windows 7.

Since *WiseMarket* is designed as an advanced tool for the enterprise-level task-holders, our empirical studies are conducted for evaluating the effectiveness and efficiency of the tool.

Synthetic datasets are generated to explore the intrinsic characteristics of the problems and algorithms; and real datasets are adopted to evaluate the usability of the algorithms. Therefore, we estimate the confidence from both ranking-based method [8] and benchmark-based method [15] on a collection of *Weibo*[1] data with 1,200 users, which could serve as a real distribution of the confidence among social online users. The Weibo service is a rapid growing platform in Chinese web community, and we choose it as a representative example for online user distribution. In addition, synthetic data sets include both normal distribution and Zipf distribution. The Zipf distribution is pervasively observed in Internet-related phenomenon, and become one of the most verified laws in the networks associated with human behaviors[1]. The basic fact of Zipf distribution is that in social networks, there is a log-linear relationship between the size and the rank of certain groups of individuals. The distribution is described by a skew-factor s, indicating how quickly the size decreases according to its rank.

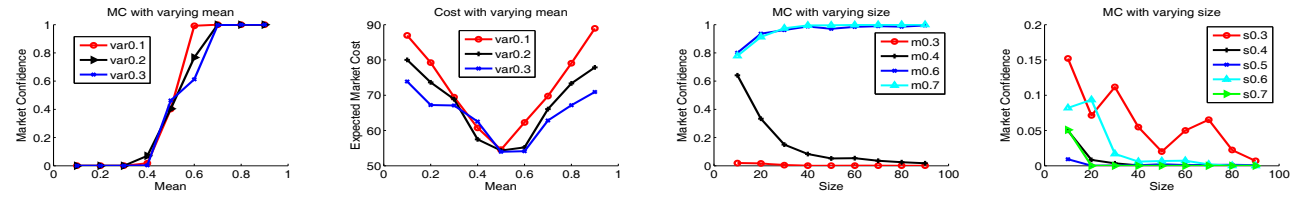
To better study the inherent computational characteristics of the EMP, we conduct an empirical study to test the intrinsic traits of MC and $Cost$, and the performance of our algorithms for calculation of MC, $Cost$, and EMP.

6.1 Intrinsic Traits of MC and $Cost$

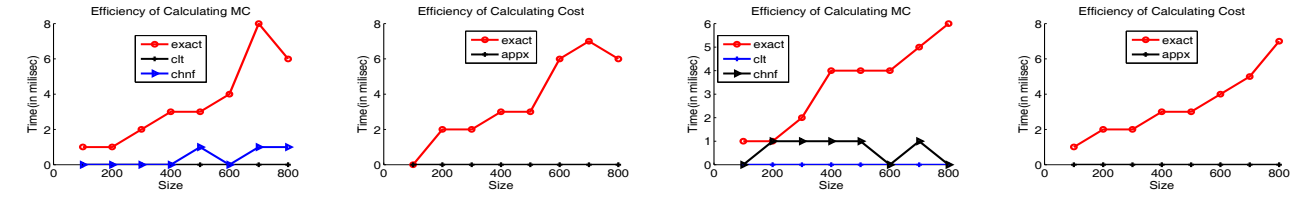
The characteristics of the *Market Confidence* and *Market Cost* over different datasets are our major interest. We generate a set of datasets following normal distribution, with varying mean value from 0.1 to 0.9 and variance from 0.1 to 0.3 respectively. Each dataset includes 1,000 candidate investors.

The results are shown in Figure 8(a) and Figure 8(b). We see that when the given mean approaches 0.5, the MC increases rapidly to 1. This indicates that a reliable *Wise Market* may be established from a relatively small group of investors. Note that when the variance is small(e.g. $var = 0.1$), the convergence is much more

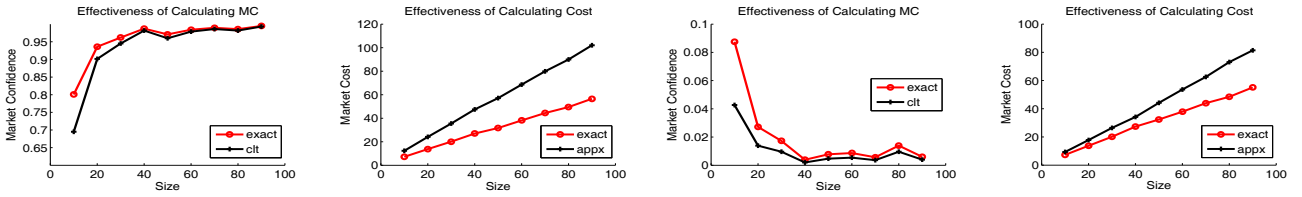
[1]www.weibo.com, the largest microblog service in China



(a) MC v.s. mean(N) (b) Cost v.s. mean(N) (c) MC v.s. size(N) (d) MC v.s. size(Z)



(e) Efficiency of Cal. MC(N) (f) Efficiency of Cal. Cost(N) (g) Efficiency of Cal. MC(Z) (h) Efficiency of Cal. Cost(Z)



(i) Effectiveness of Cal. MC(N) (j) Effectiveness of Cal. Cost(N) (k) Effectiveness of Cal. MC(Z) (l) Effectiveness of Cal. Cost(Z)

Figure 8: Experiments Results I(*N for Normal Dist., Z for Zipf Dist.)*

obvious since it is a closer approximation to Binomial Distribution. Another interesting phenomenon is in Figure 8(b), where the market with lowest Market Cost is achieved by the investors with mean of 0.5. In such *Wise Market*, the probability density distribution of the size of *Winning Set* in a *Wise Market* is close to $\lceil \frac{n}{2} \rceil$. Conceptually, either too many wise investors or too many careless investors will cause a large size of $WinningSet$, which incurs higher $Cost$ and renders the *Wise Market* ''ineffective''.

To further observe the characteristics of MC according to the change of the size of the given *Wise Market*, we run the algorithm on datasets with varying size from 10 to 90, where datasets from Normal distribution have mean from 0.3 to 0.7 and variance of 0.2, and of Zipf distribution have skew-factor s from 0.3 to 0.7. The results are shown from Figure 8(c) to Figure 8(d). In Figure 8(c), it is shown that the MC is very sensitive to the given mean value of the *Wise Market*. For the *Wise Market* whose mean value is above 0.5, the MC grows very sharply to 1 with increasing of the market size. Note that for Zipf Distribution, the MC decreases quickly to almost zero. It is because in the datasets of Zipf distribution, most of the investors have low confidence.

6.2 Performance of Algorithms for Calculating MC and $Cost$

Based on the intrinsic traits of MC and $Cost$, in this subsection we evaluate the performance, including efficiency and effectiveness, of our proposed algorithms.

Efficiency

We compare the efficiency among all proposed algorithms for calculating MC and $Cost$. The datasets have varying size from 100 to 800. In addition, for Normal distribution, the mean is set as 0.7 and the variance is set as 0.2; for Zipf distribution, the parameter s is set as 0.7.

We conduct Algorithm DC and Algorithm MCA to calculate exact MC value(denoted as *exact*), and we also conduct Algorithm based on Lemma 1 and Algorithm CLT-MC to fetch the upper bound (denoted as *chnf*) and the approximated values (denoted as *clt*). All results are shown from Figure 8(e) to 8(h). The results show great efficiency speedup by utilizing the upper bounding technique and Central Limit Theorem-based approximation algorithm. It can be observed that the time cost of exact algorithms exhibits an $\mathcal{O}(n \cdot \log^2 n)$ increasing tendency.

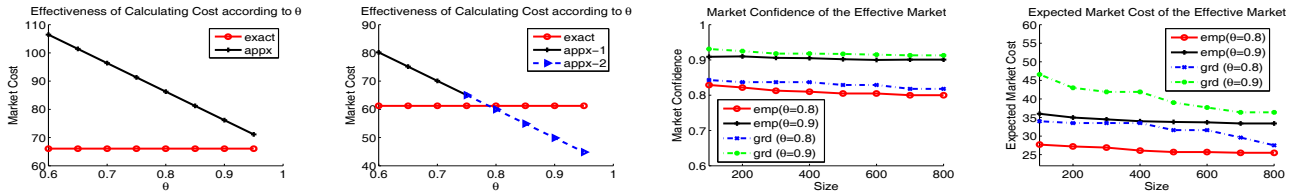
Effectiveness

The datasets are generated following a normal distribution with mean 0.6 and variance 0.2. For the Zipf distribution, the generated datasets have skew-factor s 0.2. All datasets are studied with number of nodes varying from 10 to 90. The results are shown from Figure 8(i) to Figure 9(b). The result in Figure 8(i) shows the close approximation between the exact value(''*exact*'') of MC and the one derived via CLT(''*clt*''). Due to space limitations, we present the most representative results of performance of the effectiveness issue on *Weibo* data from Figure 9(e) to Figure 9(h).

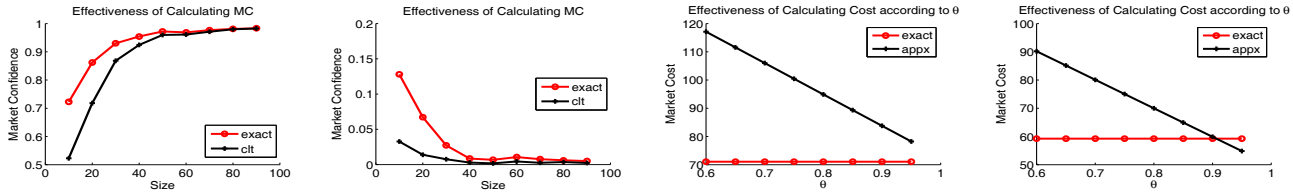
In Figure 8(j) and Figure 8(l), we observe the approximation ratio of the $Cost$ both in Normal and Zipf Distribution datasets. Instead of a converging asymptotically, the approximation of Market Cost exhibits a linear tendency(i.e.$3 - 2\theta$). Further, we vary the given threshold θ from 0.6 to 0.95, and the results of $Cost$ and its approximation are shown in Figure 9(a) and Figure9(b), where the higher the threshold, the closer the approximation is. Note that in Figure 9(b), the approximated cost becomes lower than the exact value(dashed line denoted as *appx-2*) when the given threshold is greater than 0.8. That is because the approximation only holds when the confidence of the given *Wise Market* satisfies the threshold(see Theorem 2). So, in Algorithm EMA, the approximation of $Cost$ is only considered after its MC is validated.

6.3 Performance of Algorithm for EMP

The EMP problem is essentially with an $\mathcal{O}(2^N)$ search space, so the proposed algorithm is actually an approximation algorithm.



(a) Effectiveness of Cal. Cost(N) (b) Effectiveness of Cal. Cost(Z) (c) MC of EMP (d) Cost of EMP



(e) Effectiveness of Cal. MC(B) (f) Effectiveness of Cal. MC(R) (g) Effectiveness of Cal. Cost(B) (h) Effectiveness of Cal. Cost(R)

Figure 9: Experiments Results II*(N for Normal Dist., Z for Zipf Dist.;B for Benchmark-based method, R for Ranking-based method)*

With varying size from 100 to 800 and varying threshold θ with 0.8 and 0.9, we present the Market Confidence and Cost of the selected *Effective Market*. Results are shown in Figure 9(c) and Figure 9(d). It is interesting to observe that given a threshold, our algorithm could return a *Wise Market* with a proper MC. While the size of the given candidates set increases, the MC and $Cost$ of selected *Wise Market* do not increase obviously. However, the increased threshold θ will cause a rise in $Cost$ in order to include more investors with higher individual confidence. Moreover, the dotted line represents the performance with a baseline greedy algorithm, which enrolls workers according to their confidence until the threshold is satisfied. The result shows that for the same candidate crowd, the market from our algorithm requests less market cost, which verifies our motivation for a *Wise* market.

From Figure 9(e) to Figure 9(h), we show the performance of EMP on real datasets. The results correspond to the complexity analysis in Section 5, where the approximation of MC and Cost behave satisfactory with increasing size. The difference of performance originates from different confidence estimation methods.

7. RELATED WORK AND DISCUSSION

Crowdsourcing is the human computation [19] and social computing [11] powered with crowdsourced workforce. In the context of data management, crowdsourcing is primarily used in one of two ways: 1. As a new computing power, the crowds serve as ''HPU'' that broadens the spectrum of processable data; 2. it introduces the entire social media/network as the data source instead of traditional formatted datasets. There are several leading prototype databases powered with crowdsourcing, e.g. sCoop [18], CrowdDB [9] and Qurk [14], as well as several specific operations based on these prototypes, e.g.Ranking [23], Join [13], Entity Matching [24], etc. Most of these systems and applications rely on a crowdsourcing platform like AMT as the workforce market. In addition, there are efforts to capture the wisdom of crowds from social media users, most notably in the context of opinions and reviews of popular products and travel (hotels and restaurants). Beyond these, social sensor networks are proposed to monitor earthquakes [21] and epidemics [2]. Particularly, a visionary work [6] studies decision making tasks on micro-blog service with an idealized payment model.

Another related concept is the Prediction Market, which is believed to be initially inspired by the financial markets' capability of processing miscellaneous information. In the setting of such a market, even though investors intrinsically choose the option to maximize their rewards, the choice still reflects the investor's real opinion between the two options (proven in [8]). As prototypes but also successfully running business, Iowa Electronic Market [3] and Intrade [4] are early examples of real-world prediction markets. Please also refer to [26] for a comprehensive survey of applications and models of the prediction market.

The quality of the answers from crowds is always a major concern for researchers in this area, and conducting ''workers'' selection in an active manner is one of the best practices. Related works include Active Surveying [22, 25] which studies the optimal approach to issue social inquiries. Survey sampling is a well-developed science [20]. The primary concern is to get representative samples from different demographic segments of the population, with an assumption of differences in opinion across segments. In our case, we are less interested in demographic differences -- rather we are operating in a universe with a well-understood ground truth, which we just happen not to know and hence crowdsource to learn. Moreover, online review aggregation [17, 16] has been extensively studied in the social media scenario these works typically aggregate all the inputs without any confidence or quality guarantee. Furthermore, the user contribution is driven by their own intrinsic motivation, and hence hard to control.

Discussion

Note that the crowd ''workers'' here are normal social network/media users, who spend time on the service for pleasure and entertainment. Therefore, dissimilar to the commercial crowdsourcing platform like AMT, the existence of a potential payment is already an effective arousal; besides, all the work of the users is simply making a binary choice to maximize their benefits, which also amazingly reflects their own judgement about the task (proven in [8]). As a comparison, the worker model in multi-agent systems (MAS) is designed to be more sophisticated with diverse functional attributes, which facilitates the study of logic and effectiveness of mutual cooperation. In summary, model-wise, our work provides task holders an advanced ''crowd finder'' while MAS focuses more on the worker side; and problem-wise, our work tackles a special subset problem challenge, while the MAS aims at more sophisticated optimization problems.

The limitation of this work is that we only focus on binary voting tasks, but it will not be substantially difficult to extend the setting to multiple choice problems. Moreover, this work is serving as a visionary effort in the research line of actively exploiting the power from online social users. Exciting future work includes more powerful and sophisticated tasks, where the quality of the crowds is measured differently according to the specific tasks.

8. CONCLUSION

In this paper, we define a new architecture for crowdsourcing using prediction markets defined over social media services. We define the Effective Market Problem(*EMP*) as a means for task owners to get crowdsourced answers with the smallest expected cost while meeting a specified confidence threshold.

To calculate the market confidence, we present an exact algorithm with $\mathcal{O}(n \log^2 n)$ time and a Central Limit Theorem-based approximation algorithm with $\mathcal{O}(n)$ time. To calculate the expected market cost, an exact algorithm with $\mathcal{O}(n \log^2 n)$ time complexity is provided, and an approximation algorithm, which has the $\mathcal{O}(n)$ time complexity and the $(3-2\theta)$ approximation ratio, is provided. An efficient algorithm for *EMP* by integrating the proposed algorithms. We have verified our proposed algorithms through extensive empirical studies.

We believe that this is only the first paper to exploit this new architecture, restricted to binary questions with a simple majority vote as the aggregation method. We hope, in the future work, to be able to further expand the class of questions that can be addressed fruitfully with this architecture.

Acknowledgment

This work is supported in part by the SRFDP and RGC ERG Joint Research Scheme, M-HKUST602/12, NSF grant IIS-1250880, National Grand Fundamental Research 973 Program of China under Grant 2012-CB316200, Huawei Noah's Ark Lab under Project HWLB06-15C03212/13PN, and Microsoft Research Asia Gift Grant.

9. REFERENCES

[1] L. A. Adamic and B. A. Huberman. Zipf's law and the Internet. *Glottometrics*, 3:143--150, 2002.

[2] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. EMNLP, 2011.

[3] J. Berg, R. Forsythe, F. Nelson, and T. Rietz. *Results from a Dozen Years of Election Futures Markets Research*, volume 1 of *Handbook of Experimental Economics Results*. Elsevier, 2008.

[4] J. E. Berg, F. D. Nelson, and T. A. Rietz. Prediction market accuracy in the long run. *International Journal of Forecasting*, 24, 2008.

[5] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci. Choosing the right crowd: expert finding in social networks. EDBT, 2013.

[6] C. C. Cao, J. She, Y. Tong, and L. Chen. Whom to ask? jury selection for decision making tasks on micro-blog services. VLDB, 2012.

[7] C. C. Cao, Y. Tong, L. Chen, and H. V. Jagadish. Wise market. http://www.cse.ust.hk/~caochen/tr_wm.pdf.

[8] D. A. Easley and J. M. Kleinberg. *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[9] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. SIGMOD, 2011.

[10] L. I. J. Ross, A. Zaldivar and B. Tomlinson. Who are the turkers? worker demographics in amazon mechanical turk. Technical Report SocialCode-2009-01, University of California, Irvine, USA, 2009.

[11] I. King, J. Li, and K. T. Chan. A brief survey of computational approaches in social computing. In *IJCNN*, 2009.

[12] X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang. Cdas: A crowdsourcing data analytics system. 2012.

[13] A. Marcus, E. Wu, D. Karger, S. Madden, and R. Miller. Human-powered sorts and joins. *VLDB*, 2011.

[14] A. Marcus, E. Wu, S. Madden, and R. C. Miller. Crowdsourced databases: Query processing with people. In *CIDR*, 2011.

[15] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes. We know who you followed last summer: inferring social link creation times in twitter. WWW, 2011.

[16] R. Needleman. Still more reviews aggregators: Retrevo, digitaladvisor, and thefind. http://reviews.cnet.com/, 10 2006. 19.

[17] R. Needleman. Wize: reviews aggregator tallies user feedback. http://reviews.cnet.com/, 9 2006. 20.

[18] A. Parameswaran and N. Polyzotis. Answering queries using humans, algorithms and databases. CIDR. Stanford InfoLab, 2011.

[19] A. J. Quinn and B. B. Bederson. Human computation: a survey and taxonomy of a growing field. CHI, 2011.

[20] D. Raj and P. Chandhok. *Sample Survey Theory*. Narosa Publishing House, 1998.

[21] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. WWW, 2010.

[22] H. Sharara, L. Getoor, and M. Norton. Active surveying: a probabilistic approach for identifying key opinion leaders. IJCAI'11.

[23] P. Venetis, H. Garcia-Molina, K. Huang, and N. Polyzotis. Max algorithms in crowdsourcing environments. WWW, 2012.

[24] J. Wang, T. Kraska, M. Franklin, and J. Feng. Crowder: Crowdsourcing entity resolution. VLDB. ACM, 2012.

[25] D. Watts and P. Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441--458, 2007.

[26] J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107--126, 2004.

10. APPENDIX

10.1 Correctness of Lyapunov's Central Limit Theorem

For a given set of investors WM_n, events $\zeta_1 ... \zeta_n$ are independent and may not follow a same distribution. The correctness of Lyapunov's CLT is based on the guarantee of the following two constraint conditions.

1. $E[|\zeta_i|]^{2+\delta}$ *is finite where* $\delta > 0 (i = 1, ..., n)$
2. $\lim_{n\to\infty} \frac{1}{Var^{2+\delta}(|C|)} E[|\zeta_i - E(\zeta_i)|]^{2+\delta} = 0$ *if there is* $\delta > 0$

The first condition has been proven in Theorem 1, we only discuss the correctness of the second condition here.

Firstly, for $\delta = 2$, we verify whether the second condition holds:

$$E[|\zeta_i - E(\zeta_i)|]^{2+2} = c_i(1-c_i)^4 + (1-c_i)c_i^4$$
$$= c_i(1-c_i)(c_i^3 + (1-c_i)^3) \ \ (i = 1, ..., n)$$

In addition, we can also obtain

$$Var(|C|)^{2+2} = Var^4(|C|) = (Var(|C|)^2)^2 = (\sum_{i=1}^{n} c_i(1-c_i))^2$$

According to the two formulas above, we have the following inequality.

$$0 \le \frac{1}{Var^4(|C|)} \sum_{i=1}^{n} E[|\zeta_i - E(\zeta_i)|]^4$$
$$= \frac{\sum_{i=1}^{n} c_i(1-c_i)(c_i^3 + (1-c_i)^3)}{\sum_{i=1}^{n} c_i(1-c_i))^2}$$
$$\le \frac{\sum_{i=1}^{n} (1+1)^2}{\sum_{i=1}^{n} \epsilon} = \frac{2n}{n^2\epsilon^2} = \frac{2}{n\epsilon^2}$$

So,

$$\lim_{n\to\infty} \frac{2}{n\epsilon^2} = 0$$

Based on the Squeeze Theorem, we can obtain

$$\lim_{n\to\infty} \frac{1}{Var^4(|lC|)} \sum_{i=1}^{n} E[|\zeta_i - E(\zeta_i)|]^4 = 0$$

Thus, the second condition holds. Namely, for random variables $\{\zeta_1 ... \zeta_n ...\}$, the normal sum of $\{\zeta_1 ... \zeta_n\}$ is

$$Z_n = \frac{1}{\sqrt{Var(|C|)}} (\sum_{i=1}^{n} \zeta_i - \sum_{i=1}^{n} c_i)$$

So, for any $x \in (-\infty, +\infty)$, the cumulative distribution function, $F_n(x)$, of Z_n has

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} Pr\{Z_n \le x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}} dt$$

Thus, we can know that the probability distribution of $|C|$ converges in probability to Standard Normal distribution. □