Mechanism Design for Clustering Aggregation by Selfish Systems

Pinata Winoto, Yiu-ming Cheung, Jiming Liu
Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
{pinata, ymc, jiming}@comp.hkbu.edu.hk

Abstract

We propose a market mechanism that can be implemented on clustering aggregation problem among selfish systems, which tend to lie about their correct clustering during aggregation process. Our study is the preliminary step toward the development of robust distributed data mining among selfish systems.

1. Introduction

In almost all the clustering aggregation algorithms, in order to make the clustering aggregation work, the locally obtained cluster labels are correctly reported [1, 3, 5, 6, 11, 12, 13, 14]. However, in certain scenarios in distributed data mining, several systems performing data clustering locally may not be willing to report correct labels.

In general, we are interested in the situation where systems are unwilling to share complete data but limited information. An example of this setting is sharing cluster labels of bank customers. Suppose all banks have common customers who have different relationships with each bank, which has the classification of them. However, they will not share the classification due to privacy issue or in protecting their interest. Instead, cluster labels are shared.

The interpretation of clustering aggregation results is solely on the discretion of each bank. How to implement a mechanism so that each bank will report true cluster labels and thus to convince more banks to engage in aggregation activity is the goal of our recent work. To solve it, we adopt technique which has been developed in economics for decades and in multi-agent systems recently: mechanism design.

Our analysis can be applied to selfish systems, such as e-commerce systems. Our study can also be extended to the aggregation of multiple classifier systems. The rest of the paper is organized as follows. In the next section, we will present related work. Then in Section 3 is our clustering aggregation framework, followed by mechanism design issue in Section 4. Finally, Section 5 provides future work.

2. Related work

2.1. Mechanism design

Mechanism design is a branch of game theory aiming at designing a game so that it can attain the (designer's) social objective after being played for a certain period or when it reaches an equilibrium state, assuming all players are rational. The design includes the assignment of an appropriate set of admissible strategies and payoff functions to all players.

Despite extensive studies by economists and game theorists, mechanism design has been studied in artificial intelligence community as well, especially in the context of designing multi-agent systems that can achieve a fair allocation, maximize the total utility (social welfare), and be immune from deceitful strategies [2, 9]. In computer science, mechanism design has been studied in the context of mobile ad hoc network [7], e-commerce [15], grid computing [4], etc. In this paper, we apply an ad hoc mechanism for our distributed clustering aggregation problem.

2.2. Clustering aggregation

Previous work on clustering aggregation aims for various goals, such as achieving robust results, reducing cost, protecting privacy, etc. [6, 12]. In earlier work, various clustering methods are applied to a dataset where their clustering results are then integrated to get believably more robust results [13]. In distributed database, unifying heterogeneous datasets may not be feasible due to the large size of data. Hence, a possible solution is to perform clustering locally by each node to obtain class labels that can then be integrated to get an aggregate clustering [8, 12].

In most literature researchers considered various formulations for the problems, with a major goal to ensure the quality of the final clustering result. These

1550-4786/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDM.2007.80

IEEE computer society

studies made use of different clustering mechanisms including distance measurements to form the aggregate cluster labels. For instance, Fred and Jain [5] use a single linkage approach to unify the multiple runs from the k-means algorithm performed locally; Fern and Brodley [3] use a complete linkage approach in the aggregation; Gionis et al. [6] and Johnson and Kargupta [8] also apply agglomerative algorithm in the aggregation; Topchy et al. [14] treat the clustering aggregation into a maximum likelihood estimation problem, and adopt an EM algorithm to locate the final clustering. Other methods such as dynamic programming [11], hypergraph method [12], and voting mechanism [13] are also used to improve the accuracy of clustering aggregation or to achieve scalability by reducing computational complexity.

In these prior works, the reported cluster labels are assumed to be correct and not strategically aligned. Our work complements them with respect to the distributed clustering of privacy data where systems may misreport their cluster labels.

3. Clustering aggregation framework

3.1. Model description

Consider M systems $\{s_1, s_2, ..., s_M\}$, where each of them holds a set of private attributes of the common N records $D = \{d_1, d_2, \ldots, d_N\}$. We assume crisp partition-based clustering. For a system s_m holding a set of attributes of D_m, a partition-based clustering C_m over D divides D_m into K disjoint sets, e.g. $Cluster_{m1}$, $Cluster_{m2}$, …, $Cluster_{mK}$. Suppose K is the same for all systems and $M \geq K$. Since the cluster label could be arbitrary, we denote c_{mn} to represent the cluster label given by system s_m to d_n, and for the sake of clarity, let the cluster label be integer; i.e., $c_{mn} \in \mathcal{K}$, where $\mathcal{K} \equiv \{1, 2, \ldots, K\}$. Suppose the notation $\sim c_{mn}$ refers to any cluster label other than that assigned to d_n. The cluster label reported by the system is denoted c_{mn}', where $C_m' \equiv \{c_{m1}', \ldots, c_{mN}'\}$ be the set of all reported labels by s_m. Since a system is competing with others, it may lie by reporting $c_{mn}' = \sim c_{mn}$. Also, let $l_n(k)$ be the number of systems assigning d_n to the cluster k, and $l_n(c_{mn}')$ be the total number of systems reporting d_n as the same cluster as c_{mn}' including the system m, where $\Sigma_{k\in\mathcal{K}}\, l_n(k) = M$ for any $k\in\mathcal{K}$.

Suppose system s_m assigns *confidence of clustering* values to d_n, denoted by a vector $\boldsymbol{B}_{mn}\in[0,1]^K$, with norm $|\boldsymbol{B}_{mn}| \leq 1$, which its elements, denoted by b_{mn1}, …, b_{mnK}, represent the confidence that d_n belongs to $Cluster_{m1}$ to $Cluster_{mK}$. In a crisp partitioning, the index of the greatest element in $\boldsymbol{B}_{mn}$ represents the cluster label c_{mn}. Suppose each system uses its own criteria to generate this private value. Given this, s_m may evaluate a cluster $c_{mn} = k$ in terms of utility, denote $u_{mn}(b_{mnk}, \gamma_{mnk})$, where b_{mnk} is the k-th element of $\boldsymbol{B}_{mn}$ and $\gamma_{mnk} \in \mathcal{R}_{\geq 0}$ represents factors other than confidence considered in the valuation of a clustering on record d_n, e.g. expected profit from a proper action after clustering. An intuitive yet simple utility function is:

$$u_{mn}(b_{mnk}, \gamma_{mnk}) = b_{mnk}\ \gamma_{mnk} \qquad (1)$$

Here, an accurate clustering means $b_{mnk} \rightarrow 1$, which causes $u_{mn} \rightarrow \gamma_{mnk}$. Likewise, $u_{mn} \rightarrow 0$ when $b_{mnk} \rightarrow 0$. Certainly, this utility function is not canonical to represent system preference. Rather, we use it in our analysis because of its simplicity.

3.2. Problems

In prior work, the goal of clustering aggregation problem is to find a final clustering C such that to minimize the total number of disagreements between labels in C and those in C_1', …, C_M', or to maximize common information between them [12, 13]. Here, we argue that minimizing disagreement among clusterings should not be the primary goal in our setting.

First, only if all systems believe that a minimal-disagreement clustering C^{min} approximates the "ground-truth" clustering C^*, then it can be the ultimate goal of clustering aggregation problem. This is true only if the following conditions are satisfied: (i) C^* exists; (ii) the probability of error in reported clustering is less than 0.5; and (iii) we have enough participating systems [13]. If the number of participating systems is small or some systems cheat, then $C^{min} \neq C^*$. Hence, preventing cheat and increasing participation are very important.

Second, a system shall be granted autonomy in the processing its own data and standing aside from the final consensus. In fact, after receiving cluster labels from other systems, a system s_m holds a set of private attributes of D_m plus C_1', C_2', …, C_m, …, C_M'. Since all systems hold different information, they have different interpretations. Even when all systems are truth-tellers, the correctness of a shared cluster label is still affected by the reliability of the systems in processing the data. Thus, minimizing disagreement should not be adopted as the ultimate goal. Instead, a subjective criterion should be adopted in which each system is responsible for its own aggregation and interpretation.

Instead of minimizing disagreement, we believe that the clustering aggregation problem should maximize the social welfare of participating systems, viz. $max_k \Sigma_m \Sigma_n\, u_{mn}(b_{mnk}, \gamma_{mnk})$. This approach conforms to [10], which argued that the data mining results should be valued by the decision makers.

If we only consider confidence and assume $u(x) = x$, then the problem reduces to $max_k\ \Sigma_m\Sigma_n b_{mnk}$, namely *confidence maximization problem*. If for all systems the confidence value is solely an inverse of disagreement, then our problem reduces to the minimization of disagreement. Hence, minimizing disagreement is a special case of maximizing total utility.

Since the systems are distributed and each has its own utility function $u_{mn}(\cdot)$, it is hard to maximize total utility by a central computation. Indeed, for some systems, information behind the reported cluster label is more important than the label itself. The problem that we want to solve is to find a clustering aggregation mechanism to ensure:

(i) participating systems are more reluctant to lie;
(ii) it maximizes total utility of participating systems;
(iii) it promotes the information sharing beyond the cluster labels.

To prevent lying, we design a mechanism with transferable utility (e.g. by monetary payment).

4. Mechanism design issues

4.1. Utilities, beliefs and decision structures

Suppose a system's prior confidence that d_n belongs to cluster 1 is $b_{mn1}{}^{0}$. Let $v \equiv l_n(1)$ be the number of systems that say d_n belongs to cluster 1, and $v\hat{} \equiv max_{k\in\mathcal{K}-\{1\}}(l_n(k))$ be the largest number of votes that it belongs to another cluster, assuming that the system is telling the truth. After knowing v and $v\hat{}$, the system's posterior confidence is $b_{mn1}{}^{1}(b_{mn1}{}^{0}, v, v\hat{})$ which is updated independently from its own vote. But the posterior expected profit $\gamma_{mn1}{}^{1}(\gamma_{mn1}{}^{0}, v, v\hat{})$ depends on its own vote. The utility function now is

$$u_{mn}{}^{1} = b_{mn1}{}^{1}(b_{mn1}{}^{0}, v, v\hat{})\ \gamma_{mn1}{}^{1}(\gamma_{mn1}{}^{0}, v, v\hat{}) \qquad (2)$$

Example 1. Let $b_{mn1}{}^{1} = b_{mn1}{}^{0} + (1 - b_{mn1}{}^{0})(v - v\hat{})/M$ and $\gamma_{mn1}{}^{1} = \gamma_{mn1}{}^{0}(1 - (v - v\hat{})/M)$, where M is the total number of systems. Substituting them into equation (2) yields

$$u_{mn}{}^{1}=(b_{mn1}{}^{0}+(1-b_{mn1}{}^{0})(v-v\hat{})/M)\gamma_{mn1}{}^{0}(1-(v-v\hat{})/M) \qquad (3)$$

Figure 1 depicts equation (3) for various $b_{mn1}{}^{0}\in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ and $\gamma_{mn1}{}^{0} = 100$. The x-axis is the marginal voting ratio $(v - v\hat{})/M$, where the shaded area represents negative ratios, i.e. cluster 1 does not receive the majority vote. A zero ratio means a tie between the candidate and other(s), and +1 or –1 ratio means an absolute win or loss. It is shown from the figure that an increase of the marginal voting ratio causes an increase of the utility when both the confidence and the ratio are low (area circled in Figure 1). However, when the confidence is high ($b_{ji}{}^{0} = 0.7$ or 0.8), the utility decreases as the marginal ratio increases (sharing profit with others). ■

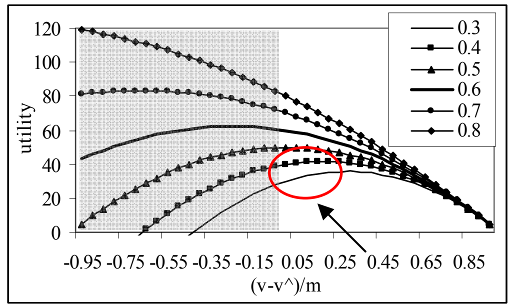

Figure 1. Utility values for various prior confidences

When a system lies, the utility only differs on the second term of equation (2), or

$$u_{mn}{}^{lie} = b_{mn1}{}^{1}(b_{mn1}, v, v\hat{})\ \gamma_{mn1}{}^{lie}(\gamma_{mn1}{}^{0}, v, v\hat{}) \qquad (4)$$

Taken equation (3) as an example, the following utility function shows the effect of lying:

$$u_{mn}{}^{lie} = (b_{mn1}{}^{0}+(1-b_{mn1}{}^{0})(v-v\hat{})/M)\gamma_{mn1}{}^{0}(1-(v-v\hat{}-a)/M) \qquad (5)$$

where $a = \{1, 2\}$ depending on whether the vote by the lying system increases $v\hat{}$ ($a = 2$) or not ($a = 1$). Nonetheless, $u_{mn}{}^{lie} > u_{mn}{}^{1}$ when $\gamma_{mn1}{}^{0} > 0$, which means all rational systems will lie when the data is profitable. However, the gain is insignificant when $a/M \to 0$ or M is large, or when $\gamma_{mn1}{}^{0} \to 0$.

In the rest of our analysis, we assume that the system will not lie for not-valuable cluster(s). Without loss of generality, suppose our analysis is for the n-th record and its cluster label is $c_{mn} = 1$, which represents a valuable cluster. Unless otherwise specified, we assume plurality voting in our setting. Lets denote $c_n{}^{\text{majority}} = 1$ when the largest group of participating systems (hereafter the *majority*) agrees that the cluster label of record n is 1 after cluster alignment. Let $q_{n1}{}^{m}$ be the belief (subjective probability) by s_m that $c_n{}^{\text{majority}} = 1$ when it reports the truth (c_{mn}' = 1), and $q_{n1}{}^{-m}$ be the same belief when it lies. Depending on the domain, we may have various relations between b_{mn1} and $q_{n1}{}^{-m}$:

- $\forall m, n\ b_{mn1} \propto q_{n1}{}^{-m}$ (*consistent belief*)
- $\forall m\ \exists n\ b_{mn1} \propto 1/q_{n1}{}^{m}$ (*partial inconsistent belief*)
- $\forall m,n\ b_{mn1} \not\simeq q_{n1}{}^{-m}$ (*uncertain belief*)

The first case will be discussed in this paper, while the second and third cases are for future work.

Suppose the minority will be punished to pay the majority the amount of y dollars. When a system s_m lies by reporting c_{mn}' $= K$, two possibilities may happen: $c_n^{\text{majority}} = 1$ by probability q_{n1}^{-m}, or $c_n^{\text{majority}} \neq 1$ by probability $(1 - q_{n1}^{-m})$. Since the system s_m is not a majority in the former case, it may be penalized to pay y, also by probability q_{n1}^{-m}; denote the utility in this case u^{C_lie}. If the majority does not choose 1, the system may receive x_l dollars as a reward, i.e. when $c_{ji}^{\text{majority}} = K$. Denote the utility in this case u^{I_lie}. Both u^{C_lie} and u^{I_lie} can be derived from equation (4) with different estimated values of v and v^. Hence, the expected utility of s_m from lying U_{lie} is

$$U_{lie} = q_{n1}^{-m}(u^{C\_lie} - y) + q_{n2}^{-m}(u^{I\_lie} - y) + \dots + q_{nK-1}^{-m}(u^{I\_lie} - y) + q_{nK}^{-m}(u^{I\_lie} + x_l) \quad (6)$$

where $q_{n1}^{-m} + q_{n2}^{-m} + \dots + q_{nK-1}^{-m} + q_{nK}^{-m} = 1$. (7)

This equation can be simplified into

$$U_{lie} = q_{n1}^{-m}(u^{C\_lie} - y) + (1 - q_{n1}^{-m})(u^{I\_lie} + x') \quad (8)$$

where

$$x' = (-q_{n2}^{-m}y - \dots - q_{nK-1}^{-m}y + q_{nK}^{-m}x_l) / (1 - q_{n1}^{-m}) \quad (9)$$

Here, x_l and y are transferable utility such as money which satisfies a *payment property*: $x_l \geq x' \geq -y$. In budget-balance mechanism, the value of x_l is not known in advanced, because we do not know the number of minority systems.

Now, when the system s_m reports the cluster label truthfully (honest), viz. $c_{mn}' = 1$, it also faces two possibilities: the majority choose cluster 1 or other label(s) by probability q_{n1}^{m} and $(1 - q_{n1}^{m})$, respectively. Denote the utility in both cases u^{C_hon} and u^{I_hon}, which can be derived from equation (2) with different estimated values of v and v^. Suppose the system receives x_h when $c_n^{\text{majority}} = 1$. The expected utility of s_m from reporting its true label is

$$U_{hon} = q_{n1}^{m}(u^{C\_hon} + x_h) + q_{n2}^{-m}(u^{I\_hon} - y) + \dots + q_{nK}^{-m}(u^{I\_hon} - y) \quad (10)$$

which can be simplified into

$$U_{hon} = q_{n1}^{m}(u^{C\_hon} + x_h) + (1 - q_{n1}^{m})(u^{I\_hon} - y) \quad (11)$$

Note, given equation (8) and (11), a rational system may not always lying, even when $u^{C\_lie} \geq u^{C\_hon}$ and $u^{I\_lie} \geq u^{I\_hon}$.

4.2. Proposed mechanism

The mechanism in Figure 2 provides a basic framework for partial truth-telling clustering aggregation among multiple selfish systems. A fully truth-telling property may be assured when we choose an extremely large y_{max}. However, this may impede the participation of systems with less confidence. One may suggest a further mechanism to decide this value, which is beyond the scope of our current work.

Step 1. All systems bid $Y_1, \dots, Y_N$ simultaneously, where Y_n is the set of preferred penalties for record n, submitted by all M systems, $|Y_n| = M$.

Step 2. All systems calculate $max(Y_1), \dots, max(Y_N)$. Any system may decide to withdraw from the mechanism after calculating these values. If not, then it will proceed to Step 3.

Step 3. All remaining systems report their clustering $C_1', C_2', \dots, C_M'$ simultaneously.

Step 4. All systems calculate C locally such that it minimizes disagreement with $C_1', C_2', \dots, C_M'$.

4.1 Permute label in $C_1', C_2', \dots, C_M'$ so that they are consistently labeled (cluster alignment).

4.2 For each record d_n, use plurality voting to determine $c_n^{majority}$; if it is tie, then leave it empty.

4.3 $C = \cup_{n \in N} \{ c_n^{majority} \}$.

Step 5. For each system s_m, if $c_{mn}' \neq c_n^{majority}$, where $c_n^{majority}$ is a non-empty value, then it pays $max(Y_n)$ which is evenly distributed to all systems z m whose $c_{zn}' = c_n^{majority}$.

Step 6. Repeat Step 1 to Step 5 until $C_1' = C_2' = \dots = C_M'$, or fewer than two systems remain in the loop. ■

Figure 2. Proposed mechanism

4.3. Analysis of the mechanism

Consider $m \geq 3$ and a system adopts utility function in equation (8) and (11). Let $u^{C\_lie} = u^{C\_hon} + \Delta^C$ and $u^{I\_lie} = u^{I\_hon} + \Delta^I$, $\Delta^C \geq 0$ and $\Delta^I \geq 0$.

Theorem 1. *The system with consistent belief will tell the truth if*

$$y > \frac{(u^{I\_hon} - u^{C\_hon})(q_{n1}^{m} - q_{n1}^{-m}) + \Delta}{M\left(\dfrac{q_{n1}^{m}}{l_n(1)} - \dfrac{q_{nK}^{m}}{l_n(K)}\right)} \quad (12)$$

where

$\Delta = q_{n1}^{-m}(u^{C\_lie} - u^{C\_hon}) + (1 - q_{n1}^{-m})(u^{I\_lie} - u^{I\_hon}) \geq 0$.

The proof of all theorems is omitted here due to limited space. From the consistent belief property $l_n(K)$ and $l_n(1)$ are directly proportional to $f(q_{nK}^{m})$ and $f(q_{n1}^{m})$, respectively. For instance, $l_n(K) \to M$ when $q_{nK}^{m} \to 1$, and decreases to M/K when $q_{nK}^{m} \to 0$. If we assume their relationship in a linear form, we have $l_n(1) = M[(1 - 1/K)\, q_{n1}^{m} + 1/K]$ and $l_n(K) = M[(1 - 1/K)\, q_{nK}^{m} + 1/K]$.

Theorem 2. *Let* $l_n(1) = M[(1 - 1/K)\, q_{n1}^{m} + 1/K]$ *and* $l_n(K) = M[(1 - 1/K)\, q_{nK}^{m} + 1/K]$, *the consistent-belief system will tell the truth if*

$$y > \frac{(u^{I\_hon} - u^{C\_hon})(q_{n1}^{m} - q_{n1}^{-m}) + \Delta}{(q_{n1}^{m} - q_{nK}^{m})K}[(K-1)q_{n1}^{m} + 1][(K-1)q_{nK}^{m} + 1] \quad (13)$$

where $q_{n1}^{m} \neq q_{nK}^{m}$ *and*
$\Delta = q_{n1}^{-m}(u^{C\_lie} - u^{C\_hon}) + (1 - q_{n1}^{-m})(u^{I\_lie} - u^{I\_hon}) \geq 0$.

Since q_{n1}^{m}, q_{n1}^{-m}, q_{nK}^{m}, u^{C_hon}, u^{I_hon}, u^{C_lie}, and u^{I_lie} are privately known, a mechanism designer can only manipulate the penalty y. Note that $[(K-1)\ q_{n1}^{m} + 1][(K-1)\ q_{nK}^{m} + 1] > 0$ because $K > 1$, and $q_{n1}^{m} > q_{nK}^{m}$ when all other systems are believed to be honest. If the system believes that other system(s) is lying, then there is a chance that $q_{n1}^{m} < q_{nK}^{m}$, i.e., the majority may vote cluster K instead of cluster 1. Hence, the necessary condition for $q_{n1}^{m} > q_{nK}^{m}$ is all other participants be believed honest, which leads to Bayesian Nash equilibrium.

Theorem 3. *Suppose for all consistent-belief systems* $l_n(1) = M[(1 - 1/K)\ q_{n1}^{m} + 1/K]$ *and* $l_n(K) = M[(1 - 1/K)\ q_{nK}^{m} + 1/K]$. *Telling the truth is the Bayesian Nash equilibrium strategy when* $u^{C\_hon} \geq u^{I\_hon} + \Delta\ (q_{n1}^{m} - q_{n1}^{-m})^{-1}$, *or* $q_{n1}^{m}\ u^{C\_hon} + (1 - q_{n1}^{m})\ u^{I\_hon} \geq q_{n1}^{-m}\ u^{C\_lie} + (1 - q_{n1}^{-m})\ u^{I\_lie}$.

Theorem 3 shows the existence of equilibrium strategy, which is very important in a mechanism design. Note that Theorem 3 holds with or without penalty ($y \geq 0$). Since $u^{C\_hon} > u^{I\_hon}$, $u^{C\_lie} > u^{I\_lie}$, $u^{C\_lie} \geq u^{C\_hon}$ and $u^{I\_lie} \geq u^{I\_hon}$, the condition would be possibly met when q_{n1}^{m} is significantly greater than q_{n1}^{-m}, or when the vote by the system counts. When the condition is not met, we need a positive penalty to ensure equilibrium as shown in Theorems 1 and 2.

From Theorems 1 and 2 we conclude that a penalty y that satisfies inequalities (12) and (13) is needed when $b_{nm1} < 1$. This penalty should be reasonable to maintain the truth-telling property, but not to discourage the participation. From an economic perspective, systems with a lower confidence about their clustering results should pay more for updating their confidence.

To find a reasonable penalty for each record, we may ask each system to bid the amount of penalty that the minority should pay within a given range, $y_{mn} \in [y_{min}, y_{max}]$. A rational system will bid y_{mn} such that the expected y_{nFinal} maximizes U_{lie} or U_{hon} whichever is the highest. If the mechanism announces $max(Y_n)$ to determine y_{nFinal}, then each system knows that $y_{max} \geq y_{nFinal} > y_{mn}$. In a special circumstance, the system may bid its *indifferent penalty* y^* that makes it indifferent between lying and telling the truth, i.e. equal to the RHS of inequality (12) or (13).

Theorem 4. *Given systems with consistent belief under Mechanism-1 with a range of allowed penalty* $[y_{min}, y_{max}]$ *and* $y_{nFinal} = max(Y_n)$, *then*

(i) Systems with linear or concave function $l_n(k)$ *bid* y_{min}.
(ii) Systems with convex function $l_n(k)$ *may bid* y_{min}, y_{max} *or any value within* $[y_{min}, y_{max}]$.
(iii) Systems with sigmoid (S-shape) function $l_n(k)$ *may bid* y^* *or any value within* $[y_{min}, y_{max}]$.

In principle, Theorem 4 shows the difficulty to elicit the distribution of y^*. If we can elicit the distribution of y^*, we may optimize Y_n so that to maximize the truth-telling of systems. A better elicitation mechanism is an open problem.

4.4. Simulation results

To analyze and visualize the relationship between confidence, maximum penalty and the effectiveness of mechanism, we have performed a simulation study. We assume seven systems using the mechanism. First, we generate a set of 100 synthetic records which may be put into two clusters. For each record, we also generate its referential c_n, b_n^0, and γ_n^0, where $c_n = 1$ (i.e. all records are valuable), $b_n^0 > 0$ and $\gamma_n^0 > 0$. Then, for each system we create its own parameters: c_{mn}, b_{mn}^0, γ_{mn}^0, v, v^, q_{n1}^{-m}, q_{n1}^{m}, and q_{n2}^{m}. c_{mn}, b_{mn}^0, and γ_{mn}^0 are generated based on their referential value, i.e. by randomly change the referential values. When the randomization is extensive, we get more heterogeneous systems. Then, v, v^, q_{n1}^{-m}, q_{n1}^{m}, and q_{n2}^{m} are generated using predetermined formula.

Our simulation consists of two parts. In Part I, we study the effect of various penalties $y_{max} = \{1, \ldots, 100\}$, where $y_{min} = 0$. We also use three groups of "true" confidence b_n^0, i.e. {low, medium, high}. In Part II, we arbitrary change the random parameters to some extreme values for stress analysis.

Figure 3 shows the results of Part I where we measure the percentage of truth-telling and accuracy against y_{max} (horizontal axis). The accuracy here refers to the correctness of the aggregated labels with respect to the referential labels c_n. All plotted data are the average value from 10 repetitions.

Two interesting results are observed. First, excessive penalty does not increase both the percentage of the truth-telling and the accuracy of clustering aggregation, as shown by an erratic but nearly flat curve when $y_{max} > 19$, which is the turning point of y_{max}. Indeed, this turning point is context dependant as we observed from the simulation results in Part II. Second, both the truth-telling and the accuracy are bounded by the system's prior confidence when y_{max} > 19, where a higher b_n^0 can help to achieve better results.

Moreover, in Part II we also observe that a low prior confidence and a high heterogeneity may reduce the accuracy to as low as 49%, which may not be acceptable for the mechanism designer. Nonetheless, our simulation has demonstrated the potential of our proposed mechanism in promoting partial truth telling in clustering aggregation among selfish systems.

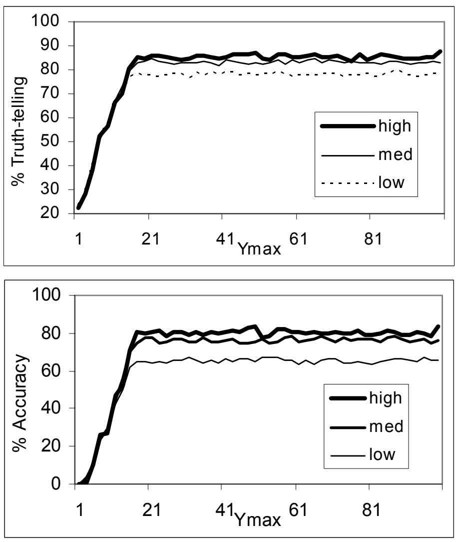

Figure 3. The percentage of truth-telling and accuracy for various b_n^0 and y_{max}

5. Conclusion and future work

In this paper we have presented a mechanism design to solve the clustering aggregation problem among selfish systems. We have applied game theory and market mechanism that are commonly studied in micro-economics to solve our problem. Although our current analysis focuses on special cases when systems have consistent belief, our approaches have opened up a new research direction to further promote distributed data mining beyond standard assumption that all systems are inherently honest. Simulation results indicate an optimal penalty may exist for a certain setting. This study can be extended by employing different voting and payment mechanisms. For example, rather than paying a flat penalty, we may set the penalty according to the proportion of the number of majority to the number of minority. Also, we may allow a negotiation on the penalty prior to the clustering aggregation. Further analysis to other cases including those with partial inconsistent belief, with soft clustering, with varying number of cluster labels, etc. are open issues. It is also interesting to perform further simulation and experiment involving (human) decision makers to find a better (partially) incentive compatible mechanism. We aim to address these issues in the future.

6. Acknowledgement

This work is partially supported by a grant from the Research Grant Council of Hong Kong Government with the project code HKBU 210306.

7. References

[1] C. Boulis, and M. Ostendorf, "Combining Multiple Clustering Systems", in *Proc. PKDD' 04*. pp. 63-74.
[2] E. Ephrati and J. Rosenschein, "The Clarke Tax as a Consensus Mechanism among Automated Agents", in *Proc. AAAI'91*, pp. 173-178.
[3] X. Z. Fern and C.E. Brodley, "Random Projection for High Dimensional Data Clustering: A cluster Ensemble Approach", in *Proc. ICML'03*, pp. 186-193.
[4] I. Foster, N.R. Jennings, and C. Kesselman, "Brain Meets Brawn: Why Grid and Agents Need Each Other", *Proc. AAMAS'04*, pp. 8-15.
[5] A.L. Fred and A. K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation", *IEEE T. Pattern Analysis and Machine Intell.*, 27(6), 2005, pp. 835-850.
[6] A. Gionis, H. Mannila and P. Tsaparas. "Clustering Aggregation", *ACM Tran. KDD*, 1(1), 2007, pp. 1-30.
[7] M. Jakobsson, J. P. Hubaux, and L. Buttyan, "A Micropayment Scheme Encouraging Collaboration in Multi-hop Cellular Networks", in *Proc. Financial Crypt. '03*.
[8] E.L. Johnson and H. Kargupta, "Collective, Hierarchical Clustering from Distributed, Heterogeneous Data", in M.J. Zaki and C.T. Ho (eds.) *Large-Scale Parallel Data Mining*, LNCS 1759, Springer, 2000, pp. 221-244.
[9] N.E. Kfir-Dahav, D. Monderer, and M. Tennenholtz, "Mechanism Design for Resource Bounded Agents", in *Proc. ICMAS'00*, pp. 309-315.
[10] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "A Microeconomic View of Data Mining", *Data Mining and Knowledge Discovery*, 2(4), 1998, pp. 311-324.
[11] T. Mielikainen, E. Terzi, and P. Tsaparas, "Aggregating Time Partitions", in *Proc. KDD'06*, pp. 347-356.
[12] A. Strehl and J. Ghosh, "Cluster Ensembles – a Knowledge Reuse Framework for Combining Multiple Partitions", *J. of Machine Learning*, 3, 2003, pp. 583-617.
[13] A.P. Topchy, M.H.C. Law, A.K. Jain, and A.L.N. Fred, "Analysis of Consensus Partition in Cluster Ensemble", in *Proc. ICDM '04*, pp. 225-232.
[14] A.P. Topchy, A.K. Jain, and W. Punch, "A Mixture Model for Clustering Ensembles", in *Proc. SDM '04*.
[15] H. Varian, "Economic Mechanism Design for Computerized Agents", in *Proc. First Usenix Workshop on Electronic Commerce*, 1995.