# Capability-Aligned Matching: Improving Quality of Games with a Purpose

Che-Liang Chiou
Department of Computer Science and Information Engineering
National Taiwan University
clchiou@gmail.com

Jane Yung-Jen Hsu
Department of Computer Science and Information Engineering
National Taiwan University
yjhsu@csie.ntu.edu.tw

## ABSTRACT

So far computer cannot satisfyingly solve many tasks that are extremely easy for human, such as image recognition or common sense reasoning. A partial solution is to delegate algorithmically difficult computation task to human, called human computation. The Game with a Purpose (GWAP), in which computational task is transformed into a game, is perhaps the most popular form of human computation. A simplified adverse selection model for output-agreement/ simultaneous-verification GWAP was built, using the ESP Game as example. The experiment results favored an adverse selection model over an moral hazard model. We were particularly interested in output quality of a GWAP affected by how players are matched with each other, and proposed capability-aligned matching (CAM) versus commonly-used random matching. The analysis showed that when compared with random mathcing, the CAM improved output quality. The experiment confirmed conclusions drawed from the analysis, and further pointed out that task-human matching scheme was as important as human-human matching scheme studied in this paper. The main contribution of this paper is the analysis and empirical evaluation of human-human matching scheme, showing that capability-aligned matching can improve quality of GWAP.

## Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

## General Terms

Economics, Experimentation

## Keywords

Game with a purpose, Adverse selection, Mechanism design

## 1. INTRODUCTION

The Game with a Purpose (GWAP) is a computer game designed to perform computation tasks as a by-product [12]. It is targeted for algorithmically difficult problems that are easy for human. Generally the GWAP are used for two purposes: (1) Solve algorithmically difficult problems. (2) Generate and/or annotate datasets for further research.

The ESP Game [11] is used as the primary example due to two reasons. (1) Because the ESP Game is the first GWAP, much of its design is widely used in many GWAP, such as output agreement, simultaneous verification, and random player matching. (2) From a game theory perspective, the ESP Game presents a fundamental type of game: static game. The analysis of the ESP Game is the basic for more complicated games, say, repeated games.

The ESP Game is used to illustrate the poor quality problem of a class of GWAPs, the output-agreement/simultaneous-verification games [12, 5].

### 1.1 Motivation

There are two orthogonal properties of outputs generated by a GWAP: correctness and quality. This paper addressed the quality of outputs. What is "correct" or "of good quality" depends on the nature of computation task; this paper defined them in the context of the ESP Game.

The ESP Game is designed to annotate images, and outputs are labels. A label is correct to an image if it describes the image. This paper defines a label is of good quality relative to another label based on their specificness. That is, labels are ordered by an "is-a" relation. For example, "red" is of better quality than "color" because red is a color but not vice-versa.

Figure 1 shows the relationship between correctness and quality. It is possible that a label is of good quality but incorrect to an image (the quadrant II). For example, "Lincoln" is of better quality than "man" but incorrect to a photo of Washington. Note: The correctness is bounded to an image, but the quality is a relative relation among labels.

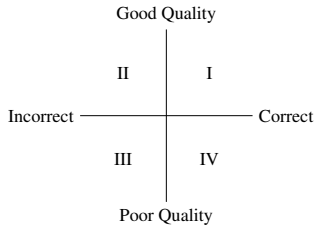


**Figure 1: The four quadrants of a label.**

In the original paper of the ESP Game [11], it has been shown that output labels are descriptions of the images, *i.e.,* correct. In a later study [10], the Google's implementation

**Cite as:** Capability-Aligned Matching: Improving Quality of Games with a Purpose, Che-Liang Chiou and Jane Yung-Jen Hsu, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 643-650.

of the ESP Game, Google Image Labeler, is examined. It is observed that players tend to answer generic labels such "building" as opposed to "terraced house", *i.e.,* correct but of poor quality.

Even worse, it has been found that output labels are predictable by a low entropy distribution. This means that a computer *without looking at the image* can guess what labels a player will output (this is exactly the cheating defined in [11]). Given that image annotation is algorithmically difficult, the predictability of outputs suggests that players are not properly motivated to outperform computers. Besides, if output labels are computer predictable, why do we need human in the course of computation?

To summarize, prior experiments have shown that output labels are correct but of poor quality (quadrant IV). Google apparently noticed this and implemented a variable scoring scheme according to specificity of labels, rather than a flat-rate scheme used in the original ESP Game. Many GWAP might also suffer from the poor quality problem because they share much of the same design of the ESP Game.

From a game theory perspective, the poor quality problem stems from the use of coordination game in the ESP Game. The focal points, also called Schelling point, of coordination GWAP are "generic label" of the ESP Game. The poor quality problem is equivalent to the existence of focal points.

The original ESP Game has implemented taboo words as an instrument to improve output quality. This is how taboo words work: When the ESP Game verifies a label, it is listed on the taboo list. Eventually all labels in the quadrant IV are in taboo list, and players can only output labels in the quadrant I—correct and of good quality.

The main shortcoming of taboo words is inefficiency: Why not simply motivate players to output quadrant I labels?

## 1.2 Overview of Proposed Solution

The *capability-align matching* (CAM) is proposed for solving the poor quality problem. Two types of of matching are in the ESP Game. (1) Matching a player with another. (2) Matching an image to a pair of players. The CAM is the former.

The CAM matches players with similar if not identical capability. On the other hand, the current implementation of the ESP Game (intentionally) use random matching.

Note: A critical game-theoretical requirement of the CAM is that which matching scheme is used is common knowledge among players.

The CAM is implemented in a small-scale experiment. The implemented method is called the *Segmentation* method, which extracts capability information from demographic data.

# 2. RELATED WORK

The ESP Game, the first GWAP, is designed to annotate images and is shown to be effective on generating label-descriptions of an image [11]. This simple game demonstrates that designing a game to use human to perform computation task is possible. Since then, many of GWAP follow its design.

Three designs of the ESP Game are relevant to modeling. They are also widely used in many GWAP.

- *Random player matching.* When compared with the CAM, it incurs poor quality output.
- *Isolated players.* This has a great strategical consequence: the ESP Game is a static game.
- *Output agreement/Simultaneous Verification.* (For its definition, see [12, 5].) It is equivalent to coordination game.

In theory, the ESP Game is a static coordination game.

The following are overview of previous approaches dealt that could be used to improve output quality.

**Incentive provision.** This approach tries to "manipulate" players through incentive [10, 6], either money or score. Its goal is to implement a designer-chosen good quality outcome. Nevertheless, in some cases incentive-provision along might be ineffective. As shown by experiments [9], increased financial incentive does not necessary increase quality. The CAM can reduce the amount of incentive required for implementing a good quality outcome.

**Competition.** This approach is based a game structure called zero-sum games. The Search War [8] is a two-player zero-sum game. The KKB [4] adds a zero-sum sub-game to the ESP Game. Nevertheless, in theory competition does not improve quality, at least in the sense of specificness, but it does diversify output—when the equilibrium is mixed. Zero-sum games often have only mixed equilibrium due to their strictly competitive nature. The zero-sum game that the Search War and the KKB use is called matching pennies, which has only a unique mixed equilibrium.

Note: This paper uses "competition" in a strict game-theoretic sense. There are GWAP [3] that are competitive (in ordinary sense) but not zero-sum. These games are coordination games, and players only compete for first proposing the will-be-agreed output.

**Community.** This approach is loosely defined by the use of social network or demographic data. It could be used for drawing players from Facebook[1], for annotating your friends [1], or for improving output quality [7]. The CAM may use communities for extract player's capability information, called the Segmentation method.

# 3. A SPECIAL THEORY OF CAPABILITY-ALIGNED MATCHING

Here we present a special version of the theory of the CAM. The general version of the theory is published in another venue due to page limits [2].

The analysis of the one-shot ESP Game is carried out by comparing the CAM in a hypothetical, ideal scenario to the random matching so that the theoretical maximal improvement of the CAM is derived. The hypothetical, ideal scenario is referred to as the first-best model, in which a computer has complete information of players' capability. The scenario of the random matching is referred to as the second-best model, in which a computer has only incomplete information of players' capability.

The performance of an outcome is defined in three aspects. These are used for comparing the first- to the second-best model.

The first aspect considers the quality, or the "revenue". It is possible that a best quality outcome is too costly to implement. The following results all condition on that there is a sufficient "margin of profit" (revenue minus cost) so that implementing a best quality outcome is in equilibrium.

[1] `http://apps.new.facebook.com/fb_gwap/`

The second aspect considers the amount of incentive provision, or the "cost". Given a particular outcome to be implemented, this amount should be as little as possible.

The last aspect considers the agreement rate, or the "risk of the business". This probability is a measure of efficiency of a game. The higher the agreement rate, the faster a correct label is produced. And why is that? The ESP Game is like a Las Vegas algorithm due to the verifying by agreement nature of it. It always produce correct labels, but its computation time varies randomly (we do not know when a pair of agents will agree).

## 3.1 The One-Shot ESP Game Model

We formulate a direct mechanism of the ESP Game, called one-shot ESP Game, in a special setting. This model is used for demonstrating the theory we will examine in experiments.

The strategic interaction in GWAP between a computer and players, in economics terminology, is a *principal-agent* relation. A computer (as a principal) hires players (as agents) to perform computation tasks. The following is the basic setup:

**Number of labels/types.** $n+1$.

**Index of labels/types.** $0 \le k, l \le n$. Note: Because $k$ and $l$ may index either label or type, we use superscript for labels, and subscript for types.

**Set of labels.** $W = \{w^k \mid 0 \le k \le n\}$.

**Qualities of labels.** $q^k = \alpha k + \beta$ for label $w^k$ where $\alpha$ and $\beta$ are real constants.

**Agent (player).** There are two agents, 1 and 2, indexed by $i \ne j \in \{1, 2\}$. The term "agent" and "player" are used interchangeably.

**Agent's output.** $w_{k,i}$ denote output label of a type-$V_k$, index $i$ agent. The index of player is often dropped because the ESP Game is symmetric, that is, $w_k$.

**Utility of agents.** $u(p) = \sqrt{p}$. Let $v = u^{-1}$ for the ease of notation. Agents are assumed to be homogeneous so that we will not be distracted from minor issues like private information of agent's utility function.

**Reservation utility** $\underline{u} > 0$. Let $\underline{v} = v(\underline{u})$ for the ease of notation.

**Capability of an agent (type).** $V_k = \{w^0, \ldots, w^k\}$. In the ESP Game, the capability of an agent is his vocabulary of words he can use. The term "capability" and "type" are used interchangeably.

**Type space.** $\mathbf{V} = \{V_k \mid 0 \le k \le n\}$. The set of capabilities, also referred to as "the type space".

**Distribution of types.** $\mu_k = \frac{1}{2^n} \mathrm{C}_k^n$ for type $V_k$. $\mu_k$ is the proportion of type-$V_k$ players. It is a binomial distribution so that most agents have moderate capability and few agents are at extreme.

**Payoffs.** The principal chooses the payoffs. In the first-best model, payoffs may contingent on both output label and type, denoted by $p(\cdot)$. In the second-best model, payoffs are contingent on output label only, denoted by $p^k$. Let $u^k = u(p^k)$ for the ease of notation.

The one-shot ESP Game is a ESP Game that a player only outputs one label. It is played as follows:

0. The quality function $q$ is given to the principal.
1. The principal chooses a payoff function $p$, and matches two agents from a pool of agents.
2. The agents observe the payoff function, and then decide whether to play (note that at this point, the agents know what matching scheme is in charge).
   (a) If any agent decides not to play, then the game terminates; the principal receives 0, and the agents both receive $\underline{u}$.
   (b) Otherwise, the game proceeds to the next step.
3. The agents simultaneously output a label $w_i$.
4. (a) If the agents agree on $w$, *i.e.,* $w = w_1 = w_2$, then the agents win; the principal receives $q(w) - p(w)$, and the agents both receive $u(p(w))$.
   (b) Otherwise, the agents lose; the principal and the agents all receive 0.

Note: For the ease of notation, the payoff to the agents is also written as

$$p(w_{k,1}, w_{l,2}) = \begin{cases} p(w) & w_{k,1} = w_{l,2} = w \\ 0 & w_{k,1} \ne w_{l,2}, \end{cases}$$

or in the unit of utility $u(w_{k,1}, w_{l,2}) = u(p(w_{k,1}, w_{l,2}))$, and the payoff to the principal

$$\pi(w_{k,1}, w_{l,2}) = \begin{cases} q(w) - p(w) & w_{k,1} = w_{l,2} = w \\ 0 & w_{k,1} \ne w_{l,2}. \end{cases}$$

A cautious reader might wonder why the payoff of the principal is not $q(w) - 2p(w)$. The reason is the ease of notation. Because only the relative order of $q$-value matters in this thesis, $q$ can be linearly scaled up arbitrarily, and whether the principal receives $q - p$ or $q - 2p$ does not matter.

## 3.2 The First-Best Model

The first-best model is our benchmark; it is the best possible performance the CAM can achieve. In the first-best model, the principal has complete information of capabilities, and players are perfectly aligned, that is, $k = l$.

The payoff function $p$ is subjected to two constraints due to the rationality of agents.

**Individual rationality.**

$$u(w_{k,1}, w_{k,2}) \ge \underline{u}. \tag{IR1}$$

**Incentive compatibility.**

$$w_{k,i} \in \arg\max_{w \in V_{k,i}} u(w, w_{k,j}). \tag{IC1}$$

The principal maximizes the average payoff

$$\max_{p, \{w_0, \ldots, w_n\}} \sum_{0 \le k \le n} \mu_k \pi(w_k, w_k) \tag{P1}$$

subjected to (IR1) and (IC1).

Obviously, the maximization program is solved by

$$p(w_k) = \begin{cases} \underline{v} & w_k = w^k \\ 0 & w_k \ne w^k. \end{cases} \tag{1}$$

That is, the agent is not paid unless he outputs a best quality label he can think of. Under this payoff function, the output label $w_k$ is, not surprisingly, the best quality output $w^k$. It is easy to check if this outcome satisfies (IR1) and (IC1). And agents always agree in equilibrium because it is a symmetric game.

So in the first-best model, the principal implements

- Best quality outcome $w_k = w^k$,
- Using minimal incentive provision $\underline{v}$,
- With perfect agreement rate equals to 1.

This too-good-to-be-true performance of first-best equilibrium shows the power of capability information in an idealize scenario. Rarely can a real world game have 100% complete capability information; there is always uncertainty in practice.

## 3.3 The Second-Best Model

In the second-best model, the principal has incomplete information; capabilities are private information of agents. The principal thus can at best randomly match agents. Note: This is usually called adverse selection in economics literature.

The Individual Rationality and Incentive Compatibility constraints are rewritten to reflect the uncertainty of the agents. Let $\Pr[w = w_{*,j}]$ denote agent $i$'s belief that agent $j$'s output is $w$

$$\Pr[w = w_{*,j}] \stackrel{\text{def}}{=} \sum_{0 \le l \le n} \mu_l \mathbf{1}\{w = w_{l,j}\}. \tag{B}$$

Note: Agent inherits uncertainty from principal, who implements random matching.

**Individual rationality.**

$$\Pr[w_k = w_*]u(w_k) \ge \underline{u}. \tag{IR2}$$

**Incentive compatibility.**

$$\Pr[w_k = w_*]u(w_k) \ge \Pr[w^l = w_*]u^l \tag{IC2}$$

where $0 \le l < k$.

**Collusion proofness.** Here is one more constraint in the second-best model than the first-best model to prevent players collude.

$$\Pr[w_k = w_*]u(w_k) \ge \Pr[w^l = w_*]u^l + \mu_k u^l \tag{CP}$$

where $0 \le l < k$.

Note: For simplicity this paper assumes that only the same type of agents can collude.

Note: Collusion is not the same as cheat defined in the original paper of the ESP Game [11] which is the attempts to fast agree on many images without looking at images. Collusion means some players lower the output quality together, but they still look at images. In other words, when players cheat, the output is incorrect (because they even not look at images). When players collude, the output is still correct, but of poor quality.

As in the first-best model, the principal maximizes its average payoff

$$\max_{p,\{w_0,\ldots,w_n\}} \sum_{0 \le k,l \le n} \mu_k \mu_l \pi(w_k, w_l) \tag{P2}$$

subjected to (IR2), (IC2) and (CP).

The three constraints are divided into two groups: (IR2) and (IC2), and (CP) alone. The best quality outcome is $w_k = w^k$, and so $\Pr[w_k = w_{*,j}] = \mu_k$. Plug them into constraint groups. The first constraint group is solved by

$$\frac{\underline{u}}{\mu_k}. \tag{2}$$

The second constraint group is solved by

$$u^{k-1} + \frac{\mu_{k-1}}{\mu_k} u^{k-1}. \tag{3}$$

The maximization program (P2) is constrained by the maximum of the two

$$u^k = \max\left\{ \frac{\underline{u}}{\mu_k}, u^{k-1} + \frac{\mu_{k-1}}{\mu_k} u^{k-1} \right\}.$$

The constraint groups are not chosen arbitrarily. They correspond to the information rent and collusion-proof rent.

So in the second-best model, the principal implements

- Best quality outcome $w_k = w^k$,
- Using amount of incentives higher than that of the first-best $u^k > \underline{u}$,
- With less than perfect agreement rate

$$\sum_{0 \le k \le n} \mu_k \Pr[w_k = w_*] < 1.$$

Here the components of second-best "cost" are analyzed, including information rent and collusion-proof rent.

**Information rent.** Observe that in the ESP Game, an agent outputs a best quality label is equivalent to an agent reveals his private information, type. Consider the first-best cost $\underline{u}$; the positive rent $\underline{u}/\mu_k - \underline{u}$ paid by the principal for acquiring agent's private information is called "information rent" in economic literature.

**Collusion-proof rent.** When $\mu$ is non-decreasing, such as when $0 \le l < k \le \lfloor n/2 \rfloor$, we have information rent inversion $\underline{u}/\mu_l > \underline{u}/\mu_k$. Does this mean a good quality label $w^k$ is paid less than a poor quality label $w^l$? In fact, no. The constraint group (3) implies that $u^k > u^{k-1}$, that is, the principal always has to pay more to a good quality label. We call this rent to maintain (CP) "collusion-proof rent".

**How much profit margin is it enough?** Now we calculate values of $\alpha, \beta$ for reference. For a best quality equilibrium to be existed, we must have positive profit margin

$$q^k - p^k > 0$$

where

$$q^k = \alpha k + \beta,$$

and

$$p^k = v\left( \max\left\{ \frac{\underline{u}}{\mu_k}, u^{k-1} + \frac{\mu_{k-1}}{\mu_k} u^{k-1} \right\} \right).$$

Let $n = 4$ and $\underline{v} = \$0.01$, then at least $\alpha \approx \$433.40$, and $\beta \approx \$2.57$. This means given that the agent's reservation utility equals to 1 cent, the qualities of labels must be worthy of tens to thousands of dollars so that a best quality outcome is still profitable after paying information rent and collusion-proof rent (see table 1). For example, the quality of label $w^4$, $q(w^4)$, must be worthy of at least \$1736.11 dollars to the principal.

On the other hand, the agreement rate is so low (roughly 27.3%) that the expected cost of the principal, the money which he actually pays, is approximately \$12.94 dollars.

This example shows us how expansive and how inefficient (in terms of agreement rate) to implement a best quality equilibrium when the principal does not have capability information.

**Example of signals.** Here we prepare results for the experiments. We considers two types of signal, "narrow" and

| $k$ | $\mu_k$ | $p^k$ |
|---|---|---|
| 0 | 6.25% | 2.56 |
| 1 | 25.00% | 4.00 |
| 2 | 37.50% | 11.11 |
| 3 | 25.00% | 69.44 |
| 4 | 6.25% | 1736.11 |

**Table 1: Payoff of labels in unit of dollar.**

"lift". Put loosely, the narrow signal is like reducing population variance, and the lift signal is like increasing population mean toward higher type. Note: Here we use the $n = 4, \underline{v} = \$0.01, \alpha = \$1000, \beta = \$10$ setup.

| $k$ | $\mu_k$ | $\mu_k \vert \theta_{\text{narrow}}$ | $\mu_k \vert \theta_{\text{lift}}$ |
|---|---|---|---|
| 0 | 6.25% | 0.00% | 5.25% |
| 1 | 25.00% | 25.00% | 25.00% |
| 2 | 37.50% | 50.00% | 37.50% |
| 3 | 25.00% | 25.00% | 25.00% |
| 4 | 6.25% | 0.00% | 7.25% |

**Table 2: The narrow and lift signal.**

Consider a signal $\theta_{\text{narrow}}$ that shrinks the population by only drawing from type-$V_1$ to type-$V_3$ (see the middle column of table 2). The expected payoff to the principal is increased by about \$217 dollars, from \$536.66 dollars to \$753.45 dollars.

Consider a signal $\theta_{\text{lift}}$ that simply add 1% to $\mu_4$ and subtract 1% from $\mu_0$ (see the rightmost column of table 2). The expected payoff to the principal is increased by about 58 cents, from \$536.66 dollars to \$537.24 dollars.

## 3.4 Summary

The principal is facing an adverse selection problem when the capability information is private, and have to pay information rent and collusion-proof rent for a good quality outcome. In addition to these rents, the principal suffers from lower agreement rate, that is, slower verification speed. We can perceive these troubles when the principal lacks the capability information as the "cost" of the random matching.

# 4. EXPERIMENT

A preliminary, small-scale experiment was conducted to test the core concepts of the theory. The experiment also demonstrated how the Segmentation method with narrow and lift signal can be implemented use online communities.

The Segmentation method extracts capability information from demographic data. The border of a demographic group is very flexible, which could be as broad as a university, or as tight as a zealous fan group. The CAM does not necessarily have to ask players to fill in annoying survey forms; the demographic data can be automatically crawled from social network websites or online forums,

Note: There is another implementation of the CAM, called the Bootstrapping method, detailed in [2].

## 4.1 Experiment Design

On choosing demographic data, the lessons will be learned from the experiments are:

- The demographic group should be related to the content of the images, and
- The deeper the participation of an agent in this group, the higher the capability he might have.

The experiment design featured:

- The one-shot ESP Game was played without any time limit.
- Subjects were not rewarded by scores or any other incentives.
- Problem sets of images that were assumed to be associated with signals were chosen.
- Subjects were actually played with robots (for reasons stated below).
- The control and treatment group differed in:
  - The matching scheme.
  - The equilibrium strategy played by robots.
- The experiment only had between-subjects effect, but no within-subjects effect (because each subject participated only once).

In brief, the experiment design features: the one-shot ESP Game, robots, and the Segmentation.

The detailed experimental process was:

1. Subjects were randomly put into either the control or treatment group.
2. Subjects were asked to report their participation level of online communities.
3. Subjects were informed the matching scheme (but actually played with a robot).

   **Control:** Random matching.
   **Treatment:** The CAM.
4. Subjects played 5 training images from each problem set, in the same order. The robot's output label was displayed when subjects lost.
5. Subjects played 20 testing images, every four from each problem set, in the same order. The robot's output label was *not* displayed when subjects lost.
6. Subjects filled in a post-hoc survey to assess the difficulty of all problem set in absolute and relative scale.

Note: In training games and test games, no scores or any other incentives were awarded when subjects won, and there was no time limit.

## 4.2 Comments on the Experiment Design

To eliminate the effect of time preference, the one-shot ESP Game, rather than the original ESP Game, was used in the experiment. The time preference should be eliminated because it has been shown to affect the output quality [6].

In addition to time preference, anything that might affect subjects was eliminated, such as time limit and scores, so that any difference in outcomes could only be explained by matching schemes.

The use of robots, a necessary evil in small-scale experiment, was because:

- It was unlikely to have aligned-capability subjects at the same time, especially when the scale was small.
- To eliminate human variations as much as possible.

The robots played equilibrium strategy, that is, pooling when put into the control group, and separating when put into the treatment group.

Only high type of robot was implemented. Otherwise, one more factor (robot's type) had to be added, along with matching scheme. This would further divide subjects, resulting in smaller groups that could not yield anything statistically significant.

In testing games, the robot's labels were not displayed to subjects when subjects lost as the original ESP Game, but in testing games, in order to teach subjects who had had never played the ESP Game, the robot's labels were displayed.

The implemented one-shot ESP Game did not compare labels literally. Instead, a list of synonyms and common misspellings was built in for label comparison.

## 4.3 Signals and Problem Sets

Two narrow-and-lift signals were chosen. One was a subject's participation level in a online community, and the other was locality (the college where subjects were recruited). Note: Although the experiment extracted these signals by asking subjects, it was easy to crawl these signals automatically.

For the online community signals, it was assumed that the population was narrower when the community was smaller, and the population was lifted higher when the participation was deeper (assuming that participation level was positively correlated to capability).

The online communities were Bulletin Board System (BBS) boards:

- WoW / Exchange information of the World of Warcraft.
- Baseball / Discussions about baseball.
- OnePiece / Discussions about the manga One Piece.

For each online community, the participation of a subject was categorized into 4 levels.

**Level 0.** None of the below.

**Level 1.** Had played the World of Warcraft, watched any of Major League Baseball games, or read the One Piece, respectively.

**Level 2.** Had read the respective BBS board.

**Level 3.** Had added the respective BBS board to his My Favorite.

Three problem sets positively associated with online community signals were chosen, namely, WoW, MLB, and OP. The MLB problem set were pictures of game characters of the World of Warcraft. The MLB problem set were pictures of Major League Baseball players. The OP problem set were pictures of manga characters of the One Piece.

Two problem sets (positively and negatively) associated with locality signal were chosen, namely, LO and FO. The LO and FO problem set were images of local and foreign celebrities and landmarks, respectively. It was assumed that subjects were more capable to the LO problem set than FO; this assumption would be verified.

The images of a problem set were carefully chosen that their difficulty to subjects was assumed uniform, and so variation of output quality within one problem set by one subject was assumed random normal.

## 4.4 Subjects

In total, 26 subjects were recruited from National Taiwan University (that means 104 labels per problem set). Table 3 shows the distribution of subject's gender, age, and group.

| | | Distribution |
|---|---|---|
| Gender | Female | 10 |
| | Male | 16 |
| Age | 18–21 | 8 |
| | 22–25 | 14 |
| | 26–29 | 2 |
| | 30–33 | 2 |
| Group | Control | 12 |
| | Treatment | 14 |

**Table 3: The distribution of subject's gender, age, and group.**

Table 4 shows the distribution of participation level. The OnePiece board had the most dedicated subjects (level 2 and level 3).

To our surprise, the WoW board was "very unpopular" among our subjects; 24 out of 26 subjects had had never played the World of Warcraft. In fact, the "unpopularity" of the World of Warcraft among subjects would cause the regression to fail because only zero or one subject was in levels above 1.

| Participation Level | BBS Board | | |
|---|---|---|---|
| | WoW | Baseball | OnePiece |
| #0 | 24 | 17 | 7 |
| #1 | 1 | 6 | 11 |
| #2 | 0 | 2 | 4 |
| #3 | 1 | 1 | 4 |

**Table 4: The participation levels of online communities.**

Table 5 shows the post-hoc survey result. The survey asked subjects to evaluate the difficulty of each problem set in absolute and relative (to other problem sets) scale.

No matter sorted by mean or median, the difficulty of problem sets were: LO (easiest), OP, FO, MLB, and WoW (hardest).

The fact that subjects felt LO was easier than FO verified our assumption that locality was a good measure of capability to the LO and FO problem set.

| | | WoW | MLB | OP | LO | FO |
|---|---|---|---|---|---|---|
| Absolute | Median | 5.00 | 5.00 | 3.00 | 2.00 | 4.00 |
| | Mean | 4.70 | 4.35 | 2.96 | 2.61 | 3.91 |
| Relative | Median | 5.00 | 4.00 | 2.00 | 2.00 | 3.00 |
| | Mean | 4.48 | 3.70 | 2.13 | 1.70 | 3.00 |

**Table 5: The absolute and relative difficulty of problem sets. The difficulty scales from 1 (easiest) to 5 (hardest).**

Table 5 also helped us verify that participation levels were indeed, as assumed to be, good measures of capability. Why was that? The participation level was an objective measure of capability, whereas the post-hoc survey was a subjective assessment of difficulty. Although different by nature, they demonstrated the same tendency: OP (most capable or easiest), MLB, and WoW (least capable or hardest) no matter

sorted by participation level or by subjective difficulty assessment.

## 4.5 Experiment Results

For each problem set, 104 labels were collected, and manually annotated. A label was annotated "of good quality" if it was the name (including synonyms and common misspellings) of a person, object, or building in the image, *i.e.,* correct and specific.

Table 6 shows numbers of label annotated as of good quality, divided by numbers of label of that category. Note: There were empty categories in the WoW problem set due to the "unpopularity" of the World of Warcraft among subjects.

A first observation was the trend that the ratio of good quality labels increased when the "Group" or "Participation Level" increased.

| Problem Set | Group | Participation Level | | | |
|---|---|---|---|---|---|
| | | #0 | #1 | #2 | #3 |
| WoW | 0 | 0/44 | 0/ 4 | 0/ 0 | 0/ 0 |
| | 1 | 2/52 | 0/ 0 | 0/ 0 | 4/ 4 |
| MLB | 0 | 6/28 | 4/12 | 0/ 4 | 2/ 4 |
| | 1 | 12/36 | 7/12 | 4/ 4 | 1/ 4 |
| OP | 0 | 0/12 | 11/28 | 0/ 4 | 0/ 4 |
| | 1 | 4/16 | 9/16 | 10/12 | 11/12 |
| LO | 0 | 30/48 | | | |
| | 1 | 44/56 | | | |
| FO | 0 | 1/48 | | | |
| | 1 | 10/56 | | | |

**Table 6: The contingency table of output labels. In the "Group" column, 0 is for the control and 1 for the treatment. Note: The LO and FO problem set did not have related participation levels.**

The ratios of good quality labels were regressed against matching scheme and participation level in a logit model. Let $i$ index over problem sets { WoW, MLB, OP, LO, FO }. Let $\Pr[Y_i = 1]$ denote the ratio of good quality labels, and $X_i$ the group (0 is for the control and 1 for the treatment), and $Z_i$ the participation level. The logit model was

$$\operatorname{logit} \Pr[Y_i = 1] = \beta_{i0} + \beta_{i1} X_i + \mathrm{DUMMY}_i \beta_{i2} Z_i \quad (4)$$

where $\mathrm{DUMMY}_i$ was a dummy variable that equaled to 1 when $i$ equaled to WoW, MLB, or OP; and 0 otherwise.

Table 7 shows the p-values of logit regressions. All were statistically significant at least at the 0.05 significance level; the null hypotheses $\beta_{i1} = \beta_{i2} = 0$ were rejected. That is, matching schemes and capabilities $(X_i, Z_i)$ indeed affected the good quality ratios $\Pr[Y_i = 1]$.

| | p-value |
|---|---|
| WoW | 0.0000*** |
| MLB | 0.0060** |
| OP | 0.0000*** |
| LO | 0.0434* |
| FO | 0.0027** |

**Table 7: The p-values of logit regressions. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.**

Table 8 shows the predictive power of the logit models; the rightmost column are p-values ($\beta$-values would be explained later). For the MLB, OP, and FO problem set, matching scheme $X_i$ and capability $Z_i$ were statistically significant predictors.

Note: The WoW and LO problem set presented interesting results. See paragraphs below.

| | | Estimate | Std. Error | z value | Pr[> \|z\|] |
|---|---|---|---|---|---|
| WoW | $\beta_{i0}$ | -26.6047 | 4380.2597 | -0.01 | 0.9952 |
| | $\beta_{i1}$ | 23.3858 | 4380.2598 | 0.01 | 0.9957 |
| | $\beta_{i2}$ | 7.7057 | 2116.2887 | 0.00 | 0.9971 |
| MLB | $\beta_{i0}$ | -1.5753 | 0.4201 | -3.75 | 0.0002*** |
| | $\beta_{i1}$ | 1.0668 | 0.4638 | 2.30 | 0.0215* |
| | $\beta_{i2}$ | 0.6083 | 0.2714 | 2.24 | 0.0250* |
| OP | $\beta_{i0}$ | -2.0770 | 0.4621 | -4.50 | 0.0000*** |
| | $\beta_{i1}$ | 1.5524 | 0.4643 | 3.34 | 0.0008*** |
| | $\beta_{i2}$ | 0.7692 | 0.2393 | 3.22 | 0.0013** |
| LO | $\beta_{i0}$ | 0.5108 | 0.2981 | 1.71 | 0.0866• |
| | $\beta_{i1}$ | 0.7885 | 0.4415 | 1.79 | 0.0741• |
| FO | $\beta_{i0}$ | -3.8501 | 1.0105 | -3.81 | 0.0001*** |
| | $\beta_{i1}$ | 2.3241 | 1.0691 | 2.17 | 0.0297* |

**Table 8: The summary of logit regression results. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.**

**WoW.** Although the logit regression failed partially due to the "unpopularity" effect (empty categories), the main reason was that the WoW problem set was the hardest. In fact, when the ratios were regressed against only capability $Z_i$,

$$\operatorname{logit} \Pr[Y_{\mathrm{WoW}} = 1] = \hat{\beta}_0 + \hat{\beta}_2 Z_{\mathrm{WoW}},$$

the p-value was statistically significant, and capability was statistically significant predictor (table 9). In other words, the WoW problem set was so hard that the capability itself dominated the outcomes, and so the matching scheme had little effect on output quality.

| | | Estimate | Std. Error | z value | Pr[> \|z\|] |
|---|---|---|---|---|---|
| WoW | $\hat{\beta}_0$ | -4.0772 | 0.7562 | -5.39 | 0.0000*** |
| | $\hat{\beta}_2$ | 2.3410 | 0.7657 | 3.06 | 0.0022** |

**Table 9: The logit regression on the WoW problem set with only capability. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.**

**LO.** On the contrary to the WoW problem set, the problem of the LO problem set was too easy. The control-LO category had good quality labels that were 30 times more than the control-FO category (table 6). Given that control group subjects should output poor quality labels (and they did in all problem sets except LO), the huge disparity between the control-LO and control-FO category may be explained by that subjects just could not think of any poor quality label; the LO problem set was just too easy.

## 4.6 Predicted Good Quality Ratios

Table 10 shows the predicted ratios $\Pr[Y_i = 1]$ from the fitted $\beta$-values. Note: The predicted ratios of the WoW and LO problem set were not predicted by statistically significant predictors, and were listed only for reference.

Consistent trends in the MLB, OP, and FO (and also WoW and LO) emerged:

- The ratios were higher when the participation levels were higher.

- The ratios of the treatment group (the CAM was used) were higher than the ratios of the control group (the random matching was used).
- The effect of matching scheme was greater than the effect of capability (participation level). Lower participation level treatment categories had higher ratios than higher participation level control categories.

The last trend was particularly interesting: A less-capable but properly-motivated player could output better quality than a more-capable but less-motivated player.

| | Group | Participation Level | | | |
|---|---|---|---|---|---|
| | | #0 | #1 | #2 | #3 |
| WoW | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0297 |
| | 1 | 0.0385 | 0.9889 | 1.0000 | 1.0000 |
| MLB | 0 | 0.1715 | 0.2755 | 0.4113 | 0.5621 |
| | 1 | 0.3755 | 0.5249 | 0.6700 | 0.7886 |
| OP | 0 | 0.1114 | 0.2129 | 0.3685 | 0.5574 |
| | 1 | 0.3718 | 0.5608 | 0.7338 | 0.8561 |
| LO | 0 | 0.6250 | | | |
| | 1 | 0.7857 | | | |
| FO | 0 | 0.0208 | | | |
| | 1 | 0.1786 | | | |

**Table 10: The predicted ratios of good quality labels. In the "Group" column, 0 is for the control and 1 for the treatment.**

## 4.7 Discussion

We had had observed:

- Potentially, a more capable player was more likely to generate good quality labels.
- The CAM had improved quality of labels, given that players had moderate capability.
- The effect of matching scheme on output quality was greater than the effect of capability, for tasks that players had moderate capability.

From the observations, a limitation of the CAM was: When difficulty of a task was extremely high or low, the capability of players dominated the output quality, and the effect of the CAM was negligible.

This limitation pointed out that matching the right task to the right player was as important as matching the right pair of players.

The experiment *per se* brought its own limitation. Subjects were interacted with robots, not other subjects, and there was only one type of robots. This experiment design restricted what we could conclude from data. The experiment was more like testing if subjects would learn and play the equilibrium strategy, and less like a user study of the CAM. Despite the methodological imperfectness, the promising results of this preliminary experiment showed that the CAM is worthy of further investigation in larger-scale experiments.

# 5. CONCLUSION

This paper proposes the capability-aligned matching (CAM) for solving the poor quality problem that the output-agreement/ simultaneous-verification Games with a Purpose (GWAP) would suffer from.

The analysis of an adverse selection model shows that the CAM has two advantages over random matching. On cost aspect, the information and collusion-proof rent, which are used for increasing output quality, are reduced. On informational aspect, the agreement rate, which is the bounding factor of verification speed, is increased.

This paper implements the Segmentation method, whose source of capability information is demographic data, and tests it in the experiments. The experiments suggest that task-human matching is as important as human-human matching.

All in all, the CAM is orthogonal to game rules, and so could be seamlessly integrated into existing and future output-agreement/simultaneous-verification GWAP for improving output quality.

# 6. REFERENCES


[1] M. Bernstein, D. Tan, G. Smith, M. Czerwinski, and E. Horvitz. Collabio: a game for annotating people within social networks. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, pages 97–100. ACM, 2009.

[2] C.-L. Chiou. Capability-aligned matching: Improving quality of games of a purpose. Master's thesis, National Taiwan University, January 2011.

[3] C.-J. Ho, T.-H. Chang, and J. Y.-J. Hsu. Photoslap: A multi-player online game for semantic annotation. In *In Twenty-Second Conf. on Artificial Intelligence*, 2007.

[4] C.-J. Ho, T.-H. Chang, J.-C. Lee, J. Y.-J. Hsu, and K.-T. Chen. Kisskissban: A competitive human computation game for image annotation. In *Proceedings of the First Human Computation Workshop (HCOMP 2009)*, June 2009.

[5] C.-J. Ho and K.-T. Chen. On formal models for social verification. In *Proceedings of the First Human Computation Workshop (HCOMP 2009)*, June 2009.

[6] S. Jain and D. C. Parkes. A game theoretic analysis of games with a purpose. In *In Proc. 4th Intl. Workshop on Internet and Network Economics*, 2008.

[7] Y.-L. Kuo, K.-Y. Chiang, C.-W. Chan, J.-C. Lee, R. Wang, E. Y.-T. Shen, and J. Y.-J. Hsu. Community-based game design: Experiments on social games for commonsense data collection. In *Proceedings of the First Human Computation Workshop (HCOMP 2009)*, June 2009.

[8] E. Law, L. von Ahn, and T. Mitchell. Search war: A game for improving web search. In *Proceedings of the First Human Computation Workshop (HCOMP 2009)*, June 2009.

[9] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proceedings of the First Human Computation Workshop (HCOMP 2009)*, June 2009.

[10] S. Robertson, M. Vojnovic, and I. Weber. Rethinking the ESP game. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 3937–3942, 2009.

[11] L. von Ahn and L. Dabbish. Labeling images with a computer games. In *In Proc. SIGCHI Conf. on Human Factors in Computing Systems*, 2004.

[12] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.