# Game Theory and Incentives in Human Computation Systems

ARPITA GHOSH
Cornell University
arpitaghosh@cornell.edu

June 13, 2013

**Abstract**

The success of a human computation system depends critically on the humans in the system actually behaving, or acting, as necessary for the system to function effectively. Since users have their own costs and benefits from participation, they will undertake desirable actions only if properly *incentivized* to do so: Indeed, while there are a vast number of human computation systems on the Web, the extent of participation and quality of contribution varies widely across systems. How can a game-theoretic approach help understand why, and provide guidance on designing systems that incentivize high participation and effort from contributors?

## 1 Introduction

The Web is increasingly centered around contributions by its users: human computation is growing increasingly common as a means for accomplishing a wide range of tasks, ranging from labeling and categorization of images and other content (with workers recruited on paid crowdsourcing platforms like Amazon Mechanical Turk, or in systems based on unpaid contribution such as Games with a Purpose or Citizen Science projects like GalaxyZoo), to answering questions on online Q&A forums (such as Y! Answers, Quora, or StackOverflow, to name a few), all the way to peer-grading homework assignments in online education. But while some human computation systems consistently attract high-quality contributions, other seemingly similar ones suffer from junk or low-quality contributions, and yet others fail due to too little participation. How can we design *incentives* in these systems to elicit desirable behavior from potential participants?

There are two components to the problem of incentive design for human computation: (i) Identifying the costs and benefits of potential contributors to the system (the components that help formulate a *model* of agent behavior), and (ii) deciding how to assign rewards, or benefits, as a function of contribution (analysis and design).

The first question of identifying costs and benefits relates closely to the question of *why* do people contribute— that is, what constitutes a benefit or a *reward*? The answer to this question, of course, varies depending on the particular system in question. While some systems (such as those based on the Amazon Mechanical Turk platform), offer financial incentives for participation, a vast majority of human computation is driven by social-psychological rewards from participation; such rewards include, for example, both intrinsic motivators like fun, interest, or the satisfaction of benefiting a cause[1], as well as extrinsic social rewards such as attention, reputation or status. There is now a growing literature in social psychology addressing what motivates, or constitutes a reward for, users in such systems[2].

But even after answering the question of why people contribute, there is a second question, which relates to how rewards are *allocated*. Given that users value rewards (by definition, and irrespective of their specific nature— financial or social-psychological), and incur costs (of time and effort) associated with different actions in the system, how rewards are assigned will influence what actions users take. That is, when a system depends on self-interested agents with their own benefits[3] and costs to participation, the quality and quantity of contributions will depend on the incentives created by the reward allocation scheme being used by the system. Given the understanding from the social psychology literature on what constitutes a reward, how should the *allocation* of these rewards be designed to incentivize desirable outcomes?

The following example illustrates the point. Consider a system with a leaderboard for top contributors (say the users who have classified the most images in a Citizen Science project like GalaxyZoo, or earned the most points in a GWAP such as the ESP game); such leaderboards appear to be strong motivators for users. While there are a number of questions related to leaderboard design, consider a very basic, simplified, question— should the system display only the top contributor, or, say, the top 5 contributors? On the one hand, if only one top-contributor 'prize' is given out, it is conceivable that users will try harder to win that solitary prize, leading to higher effort than when there are five prizes, since the presence of a greater number of prizes could mean one need not do as much to win. On the other hand, one could also argue that users will be more likely to put in effort when they know there are 5 prizes to be had, since they have a greater chance of winning something, so that their efforts are less likely to 'go to waste', in the second case where there are more prizes. Which of these is actually the correct prediction of behavior, when all participants are facing the same question of how much effort to put in? Now suppose these prizes are not positions on a leaderboard, but rather monetary rewards that all come out of a fixed prize budget (for example, as in a crowdsourcing contest)— in this case, should the entire budget be spent on one large prize or 5 smaller

[1] such as furthering science in a Citizen Science project

[2] The motivations of contributors in human computation are, naturally, closely related to those for user-generated content; some of the literature on which is discussed in [JMM12].

[3] (arising from a range of motivations including possibly other-regarding, or 'altruistic', preferences)

prizes? Again, informal arguments could be made in favor of either solution; a formal game-theoretic analysis is necessary to understand how rewards should be structured to optimally incentivize effort from contributors[4].

A formal game-theoretic approach to incentive design, very broadly, proceeds by constructing an appropriate model where users (agents) make choices over actions, which are typically associated with costs (note that the term cost does not only refer to financial costs such as an entry fee, but is also used to refer to non-monetary quantities such as the cost to time or effort). Action choices in human computation systems can consist, for example, of the following: (i) In most[5] systems, participation is a voluntary action choice (with an associated cost, e.g., of the time required to create an account or to log in to the system to participate), and mechanisms must be designed to induce adequate participation when entry is an endogenous, strategic, choice. (ii) In many systems, agents can make a choice about how much *effort* to expend on any given task, potentially influencing the quality of their output and therefore its value to the system— mechanisms must be designed so as to induce agents to expend a high level of effort (which is more 'costly' than lower effort). (iii) Finally, in some systems, agents may hold information that they can potentially strategically misreport to their benefit, such as in voting or rating— this leads to the problem of designing mechanisms that induce agents to truthfully reveal this information. (Naturally, any real system might contain a combination of these choices, as well as others unique to its function—an example of this latter kind is the choice of the order in which to output descriptive words for images in the ESP game; see §2.1).

A given design for a human computation system corresponds to, or induces, some rules that specify the allocation of rewards or benefits given each set of possible actions by agents. Note that in general, an agent's reward can depend not only on her output, but also the outputs (determined by the action choices) of other agents. Given a particular system design and the corresponding rules it induces, strategic agents will choose actions that maximize their utility (difference between benefit and cost) from the system. Agents' choices of actions lead to outputs, which in turn define the benefit, or reward, that each agent receives from the system. A vector of action choices by agents, roughly speaking, constitutes an equilibrium if no agent can improve her payoff by choosing a different action[6].

There are two aspects to a game-theoretic, or more generally, economic, approach to incentives: analysis, and design. Analyzing equilibrium behavior under the reward allocation rules of a *given* system leads to a prediction about the behavior of agents, and therefore what kind of outcomes one might expect from that system. Choosing (or altering) the rules according to which

[4]This particular problem is addressed in a model stylized for online crowdsourcing (contests, as well as crowdsourced content as in Q&A forums), in [GM12].

[5]albeit not all systems; peer-grading in online education being a prominent example

[6]A number of different *equilibrium* concepts exist to predict how strategic agents will behave under a given mechanism; see, for instance, [NRTV07].

rewards are allocated to induce agent behavior that achieves some particular outcome, or family of outcomes, constitutes *design*. While a game-theoretic approach to the analysis and design of any system with strategic agents has the general structure described above, each setting or system comes with its own unique features, depending on the choices of available actions, the nature of the available rewards and differing constraints on how they can be allocated, and *observability* of agents' outputs. In the remainder of this chapter, we will illustrate applications of the game-theoretic approach outlined above to some specific human computation domains in §2, and then discuss how the same kind of approach can be applied to reward design in the context of gamification, and rewarding contributors for their overall site participation in §3. We conclude with a discussion of challenges and directions for further work in §4.

# 2 Game-theoretic models for human computation systems

In this section, we will look at three instances of game-theoretic analysis and design for human computation systems to illustrate the game-theoretic approach outlined in the previous section. Of course, these are not the only examples of game-theoretic analysis in the context of human computation; we briefly mention two other domains of interest.

The DARPA red balloon challenge[7] was a highly publicized instance of human computation— in the sense of a distributed network of human sensors— that required incentivizing the rapid mobilization of a large number of participants on a social network. The challenge, run in December 2009, consisted of locating ten 8-foot high red balloons that had been moored at ten unknown locations throughout the US; the first team to correctly identify the locations of all ten balloons would receive a cash prize of $40,000. For a team to win the challenge, it was necessary not only to recruit members who would look for and report sightings of the balloons themselves, but also to incentivize recruits to further recruit team members, since increasing the number of searchers increased a team's chance of quickly locating the balloons. That is, in addition to the problem of incentivizing participation, a team also had to incentivize incentivizing further participation. The recursive incentive scheme used by the winning MIT team to split the prize money amongst its participants is described and analyzed in [PPR$^+$11], and highlights some of the issues that arise in the context of incentives in human computation tasks on social networks where performance, albeit not available reward, scales with the number of participants.

Another interesting family of problems related to incentives in human computation (broadly defined) occurs in online knowledge sharing or question-answer forums, such as Y! Answers, StackOverflow, or Quora, where questions posed by users are answered by other users of the site. There is a growing literature addressing a range of questions related to incentives and strategic

[7] http://archive.darpa.mil/networkchallenge/

behavior on such online Q&A forums in a game-theoretic framework, including what reward structures elicit quicker answers from users [JCP12], how to allocate attention rewards[8] amongst contributors [GM12], as well as regarding the implementability of outcomes (*i.e.*, the number and qualities of answers received) by the 'best-answer' style mechanisms used by Q&A forums such as Y! Answers [GH12].

We now proceed with an analysis of incentives and strategic behavior in three human computation settings — we discuss Games with a Purpose in §2.1, designing mechanisms for crowdsourced judgement aggregation in §2.2, and voting in the context of human computation in §2.3.

## 2.1 GWAPs

Games with a Purpose (GWAPs) [vAD08] are an outstanding family of examples of successful human computation systems. GWAPs are games designed so that people who are ostensibly simply playing the game also simultaneously produce useful input to a computation or task which cannot be performed by computers alone. For example, the game Verbosity[9] matches two players, who both 'win' if the first player correctly guesses the word being described by the second player (who, of course, is forbidden from directly using the word). This gives the second player the incentive to produce good descriptions to successfully communicate the word, thereby generating word descriptions in the process. Another game TagATune[10] pairs two players, both of whom receive a sound clip and generate descriptions for their clips to decide whether they have the same clip or not—since players 'win' when they correctly determine whether or not they have the same clip, this creates incentives for both players to generate descriptive labels for their clips, leading to a useful set of labels for sound clips in the system.

The first and perhaps best known GWAP is the ESP game[11], which cloaks the task of labeling images under the guise of a game. In the ESP game, two randomly paired players are given an image; both players are asked to generate single-word descriptions for that image. Players gain points when they agree with their partner on a descriptive word, or label, for the image (neither player can see her partner's choices until the two players have entered a common label). Since players do not know who their partner is because they are randomly paired by the game, they cannot coordinate on descriptions, and so the easiest way to agree on the output (*i.e.*, a label for the image) is to base it on the input (*i.e.*, the image itself). Thus the game design aligns the incentives of the players, who want to earn points, with that of the system, which is to generate descriptive labels for images.

But does it? The ESP game has been tremendously successful in terms of participation— it was played by over 200,000 people, collecting over 50 million tags [vAD08] in approximately 4 years since its creation. This high participa-

[8] (by choosing which answers to display, and how often or prominently to display them)
[9] http://www.gwap.com/gwap/gamesPreview/verbosity/
[10] http://www.gwap.com/gwap/gamesPreview/tagatune/
[11] http://www.gwap.com/gwap/gamesPreview/espgame/

tion makes it evident that the basic incentives were well-designed— fun was clearly a valid reward, and the game clearly generated adequate 'fun' reward to compensate for the effort involved in playing the game. But what about the *quality* of the labels generated? It has been observed, both anecdotally and in a more careful study by Weber et al [WRV08], that the labels obtained for images in the ESP game tend to have a high percentage of colors, synonyms, or generic words— essentially, labels that do not necessarily contribute too much information about the image, and are perhaps not the most useful labels that could be collected by the system. As we see next, a game-theoretic model and analysis of the ESP game can help explain how the specific choices made for the rules of the game encourage the creation of such tags, and also suggests changes to the game design which might address this issue.

Consider a simple model [JP13] for the ESP game. Each player independently chooses one of two effort levels (low or high) to exert while playing the game. A player who chooses low effort samples labels from the most 'frequent', or common, set of words in the universe (such as colors, or generic common nouns), whereas a player choosing high effort samples labels from the entire universe of words; assume that players know the relative frequencies of each word they have sampled. Next, a player can choose in what *order* to output her sampled words (which are the labels she thinks of for the image). How do the rules of the ESP game affect what effort levels players choose, and the order in which they output words?

The ESP game design rewards players as follows. Each pair of players are matched for a set of 15 images, and try to label as many images as they can achieve agreement on in 2.5 minutes. For each image, both players enter a sequence of single-word descriptions and can move on to the next image as soon as they enter a common descriptive word, which then becomes the label for the image. Players receive points for each such successful labeling. Since players can see more images (thereby potentially earning more points, since points are awarded per labeled image) if they agree quickly on a descriptive word for each individual image, the 2.5 minute time limit means that players would prefer to 'match', or agree on a label, as early as possible in their sequence of descriptive words for each image. Thus the design of the ESP game induces players to have utilities that can be described as *match-early* preferences [JP13], where each player obtains a higher utility from 'matching' earlier rather than later with her partner. What kind of player behavior, and correspondingly what kind of labels, can be expected from such 'match-early' preferences induced by the ESP game design?

**Theorem 2.1 ([JP13])** *(Informal.) With match-early preferences, choosing low effort and returning labels in decreasing order of frequency (i.e., from most common to least common) is a Bayes-Nash equilibrium in the ESP game.*

Further, it turns out that under reasonable restrictions on strategy choices, such undesirable equilibria, where players coordinate on common words, are the only Nash equilibria[12] in the ESP game. This result helps explain exactly *how*

[12] A Nash equilibrium is a set of *strategies*, one for each player, such that no player can

the design choices, *i.e.*, the specific rules of the ESP game, might lead to the observed outcomes of common or generic labels for images.

Now suppose rewards are instead designed so that the number of points received by a pair of players depends not just on the *number* of matches, but also on the *quality* of each match, based on the frequency of the agreed-upon label. Such a reward scheme, where a player's utility depends not on *when* the match occurs (*i.e.*, at which point in the sequence of words output by the player), but rather on the frequency of the matched label, induces *rare-words* preferences. How does changing the reward structure to remove the 'need for speed', and so that agreeing on rare labels leads to higher rewards, affect equilibrium outcomes?

**Theorem 2.2 ([JP13])** *(Informal.) With rare-words preferences, returning labels in decreasing order of frequency (i.e., common words first) is a strictly* dominated[13] *strategy. Returning words in* increasing *order of frequency (i.e., least common words first) is an ex-post Nash equilibrium in the ESP game, conditional on both players choosing the same level of effort.*

That is, such a change in the reward design leads players to 'try' the rarer words in their sample first, leading to more useful labels than those obtained under the equilibrium strategy of trying more common words first under match-early preferences. This change in design alone, though, is not adequate to induce effort— high effort sampling need not be an equilibrium strategy in the ESP game even when rewards are modified to induce rare-words preferences. If, however, the distribution of words in the dictionary from which samples are drawn is Zipfian (as is the case for the English language), and if the rewards are designed so that utilities additionally obey a certain (multiplicative or additive) structure, high effort sampling followed by coordination on rare words now becomes an equilibrium in the game.

This analysis of the ESP game demonstrates both (i) how a game-theoretic model and analysis can explain and pinpoint in what way a particular design choice for the game leads to the observed outcomes of low-information labels (arising from coordination on common words), and (ii) what kind of reward redesign can lead, under what conditions, to high-effort coordination on rare words. In the next subsection, we investigate another family of human computation systems where a formal analysis of incentives can aid the design of reward mechanisms that induce desirable behavior from participants in the system.

## 2.2 Crowdsourced Judgement Elicitation

An increasingly prevalent application of human computation is in the domain of using the crowd to make evaluations, or *judgements*. Suppose each of a set of objects has one of many possible properties or belongs to one of many categories,

benefit by deviating from her strategy given the strategy choices of other players; see, for instance, [NRTV07].

[13] A strategy is strictly dominated if there is another strategy that always leads to larger payoffs regardless of other players' choices, *i.e.*, for all possible strategies of other players.

and the task is to judge, or evaluate, what property the object has or which category it belongs to— for instance, categorizing galaxies or identifying birds (as in Citizen Science projects), deciding whether some text content is abusive or an image is pornographic, or deciding whether a homework assignment is correct or incorrect, or what score it should get. When the number of objects to be evaluated is too large for a single expert and the evaluation cannot be accurately performed by a computer, a human computation-based solution is to replace the expert's opinion by an aggregate evaluation based on judgements from a 'crowd' of non-experts, typically recruited via some online platform. Crowdsourced judgement elicitation is now used in a wide range of applications including image classification, identifying adult content online, rating learners' translations on the language-learning site Duolingo, and most recently for peer grading in online education, where Massively Open Online Courses (MOOCs) with huge enrollments crowdsource the problem of evaluating homework assignments back to the students in the class.

Consider a worker, say, on Amazon Mechanical Turk who is classifying images, or a Duolingo user who has been asked to rate another user's translation into his native language. Such a worker could potentially just arbitrarily categorize the object (an image, a translation, and so on) into some category— incurring no effort cost, or alternately, she can put in effort to properly evaluate the object. *If* the system could check the accuracy of the worker's output (e.g., the correctness of her categorization), and reward based on accuracy, the worker might be incentivized to put in effort into making judgements more accurately— but the reason for using human computation, of course, is that the system does not have this information in the first place. Given that the only source of information about the ground truth— the true category for each object— is judgements from the crowd, how should the system reward agents based on the received reports?

This question is related, although not the same as, the growing literature on mechanisms for *information elicitation*, also pertinent to human computation. Broadly, that literature addresses the question of designing mechanisms that incentivize agents to *truthfully* reveal information they already happen to possess, such as their opinions about a product or service (as in the peer-prediction literature [MRZ05]), or their beliefs about the probabilities of an event (as in prediction markets, a literature by now too vast to properly discuss here (Chp. 26, [NRTV07]). The problem encountered in the crowdsourced judgement elicitation domain is somewhat different than the one addressed by this literature, since here agents (workers) do not already possess the information they are being asked to share— they must expend an *effort cost* to acquire that information in the first place. Of course, having acquired the information, the reward structure additionally needs to induce agents to truthfully report what they observe.

Given both formal studies [IPW10] and anecdotal reports[14] of effort-shirking by raters under ad-hoc or output-independent reward structures in real-world systems, there is a need for mechanisms that will incentivize agents to exert

[14]such as in Duolingo and peer-grading systems

effort to make useful judgements on their tasks. Suppose an agent's utility is the difference between the reward she receives, and the cost of the effort she puts in, aggregated over all the tasks she performs. A mechanism for judgement elicitation in such human computation settings should make it 'most beneficial', if not the only beneficial strategy, for agents to not just *report* their observations truthfully, but to also to expend effort to *make* the best observations they can in the first place, rather than simply making arbitrary reports. Also, it is even more important here to ensure that the payoffs from an outcome where all agents blindly and consistently report the same observation (such as declaring all content to be good) are strictly smaller than the payoffs from truthfully reporting observations of the actual input, since declaring all tasks to be of some predecided type (without even observing the input) requires no effort and therefore incurs no cost, whereas actually putting in effort to make observations about the input will incur a nonzero cost. [DG13] provide a simple model for this setting of crowdsourced judgement elicitation with unobservable ground truth, where an agent's proficiency— the probability with which she correctly evaluates the underlying ground truth (*i.e.*, the true category or property of the object)— is determined by her *strategic choice* of how much effort to put into the task. They provide a mechanism— a set of rules which determines how to allocate rewards to agents— $\mathcal{M}$ , for binary information elicitation for multiple tasks when agents have such endogenous (*i.e.*, strategically determined) proficiencies, that has the following properties.

**Theorem 2.3 ([DG13])** *Exerting maximum effort into making judgements, followed by truthful reporting of observations is a Nash equilibrium in mechanism $\mathcal{M}$. Further, this is the equilibrium with* maximum payoff *to all agents, even when agents have different maximum proficiencies, can use mixed strategies, and can choose a different strategy for each of their tasks.*

Informally, the main idea behind the mechanism $\mathcal{M}$ is to use the presence of *multiple* tasks and ratings to estimate a reporting statistic that identifies and penalizes *blind*, or low-effort, agreement— since the only source of information about the ground truth comes from agents' reports, it is natural to use agreement as a proxy for accuracy, and reward an agent for agreement with another agent's evaluation of the same task. However, rewarding only for agreement can lead to low-effort equilibria with high payoffs (for instance, where all agents report the same observation independent of the input and therefore always agree), which is undesirable. The mechanism $\mathcal{M}$ therefore does reward agents for agreeing with another 'reference' report on the same task, but also penalizes for *blind agreement* by subtracting out a statistic term, which is based on the extent of the agreement that would be 'expected anyway' given these agents' reports over all the other tasks they rate. This statistic term is designed so that agents obtain nonzero rewards *only* when they put in effort into their observations, and so that reward is increasing in effort: this yields the maximum payoff property of the full effort-truthful reporting Nash equilibrium.

This crowdsourced judgement setting thus demonstrates another instance in which game-theoretic models and mechanism design provide useful input into

the incentive-centric design of a broad family of human computation systems, where— given the accounts of effort shirking by raters under ad-hoc or output-independent reward structures in real-world systems— properly incentivizing agents is key to obtaining worthwhile, or valuable, input from the humans in the system.

## 2.3 Aggregating quality estimates: Voting

We illustrate a third kind of incentive problem in human computation by examining settings where user ratings are used to compute the (absolute or relative) quality of online content, such as photographs on Flickr, reviews on Amazon or Yelp, shared articles on Reddit, and so on. Rating and ranking are natural applications for human computation— in all the examples we just mentioned, it is hard for a computer to accurately process the task at hand, which is inferring content quality or rankings (for example, how does Flickr know whether a photograph is appealing?), whereas humans can easily accomplish the task.

Where do incentives and game theory come in? In a number of such voting or rating contexts, the set of people producing ratings is not disjoint from, and often has high overlap with, the set of people producing the content or objects[15] to be rated (for example, consider a community of photographers such as on Flickr, who both post photos themselves, and rate other contributors' photos). Since having a high relative rating for one's own content is desirable (highly-ranked content receives more attention, which seems to be clearly desired by contributors), a contributor who is rating other contributions might have an incentive to strategize her votes so as to increase her relative ranking— for instance, by downvoting other highly-rated contenders. A natural question then is the following: Is it possible to design a scheme for aggregating ratings that can 'get at' the true qualities, or perhaps the true underlying ranking of objects, or identify the set of the $k$-best objects, when the creators of the objects being rated are also the raters?

A simple abstract model for this problem is studied in [AFPT11]. Suppose, for simplicity, that the set of raters is exactly the same as the set of creators of the content; abstractly, this can be modeled by a voting scenario where the set of agents who vote are identical to the set of candidates being voted on[16]. Consider a directed graph over this set of $n$ agents, where an edge from agent $i$ to agent $j$ is taken to mean that $i$ 'upvotes' or supports (for example, likes the content produced by) agent $j$[17]. Suppose the system wants to find the $k$ most popular agents— for example, a site might want to prominently display the $k$ most popular contributions. Each agent is only interested in being selected in this set of $k$ 'winners', and so may misreport its opinions, or ratings, to this end. A

[15] Note that these objects can also be the producers themselves, rather than only the content produced, as might be the case when constructing rankings of users based on their contributions in some online community.

[16] An example of such a situation, outside of the context of human computation or the Internet, is the election of the pope in the papal conclave.

[17] For readers familiar with the voting literature, this setting is a special case of *approval voting* where the set of voters coincides with the set of options.

*mechanism* in this setting is a way to aggregate the set of votes from the $n$ agents into a set of $k$ selected agents. Is it possible to design a mechanism which is simultaneously *strategyproof*— *i.e.*, where no agent can benefit by misreporting which other agents she approves (or does not approve) of, *i.e.*, her edges— as well as *approximately optimal*, in the sense that the total number of votes on the chosen set of $k$ agents is 'close' to (*i.e.*, not much smaller than) the total votes for the $k$ most popular agents? [AFPT11] analyze strategic behavior in this model to first show a surprising impossibility result:

**Theorem 2.4 ([AFPT11])** *For any number of agents $n \geq 2$, and any number of winners $k$ between 1 and $n-1$, there is no* deterministic *strategyproof $k$-selection mechanism with a finite approximation ratio.*

However, [AFPT11] constructs a *randomized* mechanism (*i.e.*, where the choice of the set of $k$ winners also depends on the outcome of some random coin tosses) which is both strategyproof, and selects a reasonable set of agents:

**Theorem 2.5 ([AFPT11])** *For any $k$ between 1 and $n-1$, there is a* randomized *$k$-selection mechanism that is both strategyproof, and has an approximation ratio*[18] *no worse than 4; this mechanism is approximately optimal as $k$ diverges.*

Together, these results, based on a formal analysis of strategic behavior in a simple voting model, establish the tradeoffs that the designer of a human computation-based rating or ranking system should expect to find when dealing with self-interested users— while no simple (*i.e.*, deterministic) mechanism for aggregating ratings can be both strategyproof and optimal for all inputs, there exists a more complex (randomized) mechanism that can eliminate any benefits from misreporting while also not compromising the quality of the winner set too much, especially as the size of that set diverges.

# 3 Incentivizing consistent effort: Gamification and game theory

In the previous section, we saw the role of formal game-theoretic analysis and design in three human computation contexts— specifically, we saw how rewards, or benefits, for particular tasks can be restructured to provide incentives to agents to undertake the 'right', *i.e.*, system-desired, behaviors. In this section, we will discuss an application of game-theoretic techniques to a broader class of incentives for participation: an increasing number of human computation systems are now accompanied by corresponding *online communities*, with discussion forums, leaderboards, reputation scores, and various other features, all of which also provide rewards (typically of a social-psychological nature) to participants, albeit not for performance on a particular task. While our previous analyses looked at incentives and cost-benefit tradeoffs from a *single* action or contribution, there are also rewards that relate directly to the identity of a

[18] That is, the set of winners obtains at least 1/4 as many votes as the $k$ most popular agents

*contributor* typically based on her overall contribution, rather than to single actions or contributions. In this section, we will discuss very recent work on formal approaches to designing incentives that motivate *overall contribution* in human computation systems via their communities[19].

A common theme in a growing number of online communities and social media sites relying on user contributions is *gamification*— via badges, leaderboards, and other such forms of (competition or accomplishment based) social-psychological rewards. These rewards, meant to provide an incentive for participation and effort on a given system or site, usually reflect various site-level accomplishments based on a user's cumulative 'performance' over multiple contributions. Such badges or top-contributor lists clearly appear to motivate users, who actively pursue and compete for them—for example, users on StackOverflow are observed to increase their effort levels when they get close to the contribution level required for a badge [AHKL13], and there are entire discussion communities on the Web centered around how to break into Amazon's Top Reviewer list or how to maintain a Top Contributor badge on Yahoo! Answers, while users who have just earned entry into top contributor lists often find an increased number of negative votes from other users attempting to displace them.

Given that the rewards created by these virtual badges and leaderboards appear to be valued by users (a phenomenon that appears to be quite general, occuring across a range of online communities) and that participating and putting in the effort required to obtain them is costly, a particular way of allocating these rewards creates a corresponding set of incentives, or more formally, induces a *mechanism* in the presence of self-interested contributors. So gamification also involves reasoning about incentives in a game-theoretic sense— given that there are several different ways to 'gamify' a site, how should these rewards for overall contribution be designed to incentivize desired levels of contribution? For instance:

1. What incentives are created by mechanisms induced by an *absolute* standard of output that must be met to earn a badge (such as a threshold number of images that must be tagged, or questions that must be answered), and what incentives are created by a *competitive*, or relative, standard, such as top-contributor badges or leaderboards? And how do these 'compare'?

2. When badges are awarded for meeting absolute standards, should multiple badges be awarded, and if yes, how should they be 'placed' relative to each other in terms of the accomplishments required to earn successively higher levels of badges?

3. Consider a very simple form of a relative standard, corresponding to handing out an (identical) 'top-contributor badge' to some set of 'best' contributors on the site. How exactly should badges for competitive standards be specified—should the site award some fixed number of top-contributor

[19] For a broad set of general guidelines on incentivizing participation and engagement in online communities, see [KRK+12].

badges *independent* of the number of actual participants, such as a Top 10 Contributors list (call this mechanism $\mathcal{M}_\rho^p$), or should the number of winners be some fraction of the number of *actual* participants (mechanism $\mathcal{M}_\rho^c$)? Note that since participation in all these human computation systems is a voluntary choice, the number of actual contributors is *not fixed* apriori, but rather is determined by the choices made by self-interested users— so these two specifications are *not* equivalent.

This family of questions brings us to the frontiers of research on game theory for human computation, which we summarize below. First we address the questions about what kinds of incentives are created by absolute and relative standards mechanisms. Call the awarding of badges for achieving some absolute standard, say $\alpha$, of output (such as receiving $\alpha$ positive ratings for one's contributions, or labeling $\alpha$ images correctly), an absolute standards mechanism $\mathcal{M}_\alpha$. Call the awarding of badges for belonging amongst some set of top $\rho$ contributors to the site a relative standards mechanism $\mathcal{M}_\rho$. [EG13] investigates the existence and nature of equilibrium outcomes in these two classes of mechanisms in a simple game-theoretic model where users who value badges (presumably for social-psychological reasons), and have a cost to effort, strategically choose whether to participate and how much effort to put into the site[20].

[EG13] find that even the existence of equilibria for relative standards mechanisms $\mathcal{M}_\rho$ depends on *how* the number of top contributor awards $\rho$ is specified (*i.e.*, whether there are a fixed number of top-contributor badges that will be awarded, or whether the number of badges scales as a fraction of the number of actual participants)— this is due to endogenous participation, *i.e.*, the fact that users make a voluntary choice about whether to participate depending on the rewards being offered. While the two versions of the relative standards mechanism behave identically for $\rho$ lying in a certain range, the result below suggests that at least for settings that are reasonably captured by the model in [EG13], the mechanism corresponding to announcing a fixed number of top-contributor badges that is independent of the number of actual participants is a more robust mechanism than one that declares some fraction of participants to be winners, *i.e.*, where the number of winners scales with the number of actual contestants.

**Theorem 3.1 ([EG13])** *(Informal.)*

1. *For relative standards mechanisms $\mathcal{M}_\rho$, equilibria exist for all values of $\rho > 0$ if the site specifies $\rho$ as a fraction of* potential *contributors, i.e., as a* fixed number *of winners, but not if $\rho$ refers to a* fraction *of* actual *contributors.*

2. *For absolute standards mechanisms $\mathcal{M}_\alpha$, equilibria exist for all possible values of the standard $\alpha$. However, there is a maximum standard $\alpha_{\max}$ such that the only equilibria for all standards higher than $\alpha_{\max}$ involve zero participation, leading to no contributions.*

[20] An equilibrium here consists of some level of participation and some level of effort from participants, such that no participant can benefit from either dropping out or choosing to exert a different level of effort, and no non-participant would prefer to participate.

This equilibrium analysis suggests an interesting contrast between using relative and absolute standards for rewarding overall contribution— while $\mathcal{M}_\rho^p$ elicits non-zero participation in equilibrium for every value of $\rho > 0$, $\mathcal{M}_\alpha$ can lead to zero equilibrium participation when $\alpha$ is too large. However, there is also a *partial* equivalence between absolute and relative standards $\mathcal{M}_\alpha$ and $\mathcal{M}_\rho^p$, of the following form. Every absolute standard $\alpha \leq \alpha_{\max}$ leads to an equilibrium outcome that is identical, in terms of induced effort and participation, to the equilibrium outcome in the relative standards mechanism with some appropriate value of $\rho \in [\rho_{\min}, 1)$, where $\rho_{\min} > 0$ is the equilibrium fraction of winners at the standard $\alpha_{\max}$— and in fact, the value of $\rho$ that elicits the *maximum* effort from contributors occurs at a relative standard $\rho$ that lies in this range $[\rho_{\min}, 1)$. So for a site designer who wants to optimize elicited effort, and has adequate information about the parameters of the population to choose an optimal value of the standard $\alpha$ or $\rho$, the absolute and relative standards mechanisms are equivalent. In the absence of such information, however, or with uncertainty about the population's parameters, a 'top contributor' style mechanism $\mathcal{M}_\rho^p$ based on competitive standards that always elicits non-zero equilibrium participation might be, informally speaking, more desirable than an absolute standards mechanism.

Finally, we ask a question about multiple badges— consider badges that are handed out for absolute achievements. At what levels of achievement should badges should be awarded to to sustain effort on the site, and how should they be designed to steer user behavior towards different actions on the site? [AHKL13] address this question in a model where there is a multi-dimensional space representing the possible types of actions on the site. Users have a time-discounted value to earning badges and incur a cost when they choose actions from a distribution that differs from their preferred mixture of actions on the site. If users act to maximize their utility in this model of costs and benefits, how should badges be placed to align ideal user behavior with users' utility-maximizing actions? [AHKL13] finds that the effectiveness of badges in inducing desirable behavior depends significantly on their 'placement' (*i.e.*, for what level of contribution they are awarded), with the optimal location being, roughly speaking, one that is hard to achieve and therefore motivates users for a significant length of 'time' (contributions). Also, multiple badges should be 'spread out' with roughly equal values, rather than placing them at nearby levels of contribution, suggesting that multiple smaller rewards provide more effective incentives than a small number of larger rewards at least in settings that are well-described by the model in [AHKL13].

The literature on a game-theoretic approach to overall contributor reward design is very young, and has looked at the most immediate questions under relatively simple models and reward structures. There are a number of questions still to be modeled and answered, an immediate one being the design of leaderboards. In contrast to top-contributor badges, not all 'winners' receive equal rewards in leaderboards since arguably, the reward from placing first (or say in the top 5 positions) is somewhat larger than, say, ranking 100th on the leaderboard, even in a site with a large population. A number of interesting

game-theoretic questions arise, starting from the very basic question of how many positions the leaderboard should have to optimally elicit effort from contributors; this question is related to our motivating example early in this chapter, and a first step towards such questions, although in a model with perfectly observable outputs, is taken in [GM12].

Finally, a commonly used reward structure is that of user reputations. The question of how to design—and use and update— user reputations to create the right incentives in a human computation system is one that can draw from a vast body of literature on the design on reputation systems (Chp 27, [NRTV07]), but comes with challenges unique to human computation systems that will require the development of convincing new models and schemes [21]: In addition to differences in details from the models in prior work on reputation systems (for example, in the context of electronic marketplaces such as EBay or Amazon), there are also potentially fundamental differences that might arise due to the differences in the nature of the rewards that agents seek from these systems, which are primarily financial in online marketplaces but to a large degree social-psychological (such as status or reputation within a community) in human computation systems. We briefly explore these ideas in §4.

# 4 Challenges and further directions

In the previous sections, we saw how a game-theoretic, or more broadly, an economic approach, can help with analyzing strategic behavior and incentive design in human computation systems. But there remain many challenges, unique to such online contribution domains, that need to be understood before we can fully develop the game-theoretic foundations for incentives in human computation. First, of course, there are a number of immediate questions regarding theoretical modeling and analysis. In addition to questions we have already alluded to in previous sections, there is also an interesting family of problems arising from the diversity of roles that participants play in many systems (for example, contribution versus moderation in an online community). How should incentives be designed to ensure that each participant is incentivized to properly contribute to her role(s) in the system, given that different roles might require different incentives, and that these incentives could potentially interact with each other? A principled framework that helps answer this question will need to begin with new models that appropriately capture such multi-role participation as well as interactions between different sets of incentives— an issue relates, at least in spirit, to the question of what incentives are created by simultaneously using different forms of gamification on a site. A further question along these lines, arising from the voluntary nature of participation, is how to structure incentives to also induce different potential participants to *choose* their socially optimal roles in the system.

In addition to problems related to modeling and theoretical analysis, there are also a number of cross-disciplinary questions. One family of problems lies

[21] For preliminary work on social norms for reputation, see [HZVvdS12].

at the interface of game theory and *interaction design*. By influencing usability, and usage, the design of the user interface in a human computation system also interacts with incentives in a game-theoretic sense— after all, any game-theoretic analysis involves modeling the behavior of the agents (*i.e.*, users) in the system, which is determined not only by its rules for reward allocation but also by its interface. As a very simple example, consider a system that rewards contributors based on the quality of their outputs, as measured by the ratings, or votes, provided by users who view these contributions. An interface design which leads to very little rating by users (for example, a hard-to-find rating button or an overly complex menu of options), or one that leads to ambiguity in the meaning of a rating (such as a thumbs-up button which is interpreted by some users to mean 'Helpful' and others to mean 'I agree') results in 'noisier' ratings than an interface which elicits meaningful votes from a large number of users. A greater degree of noise, roughly speaking, means that reward depends on effort in a more uncertain way, which in turn affects the incentives for agents to put in effort in the system. It is easy to see that even in this specific example there is much more to consider at the interface of interaction design and incentives, such as the question of *which* users are allowed to rate contributions, and whether raters are offered a more or less expressive set of ratings to choose from. Another example of the connection between interaction design and game theory can be found in the context of badges and gamification— how much information about users' behavior and performance is revealed to other users can potentially affect users' valuations of badges, and consequently their strategic choices; see §5.3 in [EG13]. Generally, therefore, how users respond to a given mechanism in a strategic or game-theoretic sense, as well as the space of available mechanisms itself, can depend on the choice of interface in the interaction design phase— an ideal design paradigm would take into account both the influence of the user interface and the reward allocation rules on user behavior to provide an integrated, complete approach to the design of incentives in human computation systems.

Finally, a very important family of questions relate to properly understanding contributor motivations and rewards in a more nuanced fashion. One particularly interesting issue that is pertinent to most human computation systems is that of *mixed incentives*: unlike in most traditional economic analysis, human computation systems typically involve a *mixture* of potential contributor rewards. Systems with financial rewards for contributing, such as Amazon Mechanical Turk, mix two entirely different kinds of rewards (financial and social-psychological); even in systems without financial incentives, there are usually multiple social-psychological rewards, either intrinsic or site-created: for instance, [vAD08] describes fun as the primary motivator in the ESP game, but there are also social-psychological rewards from leaderboards (competition) as well as from successful 'collaboration' with partners on the image labeling task.

How do people— the agents in a game-theoretic model— value these different kinds of rewards in combination, and also, how do they value them relative to each other? What happens when virtual points are used to create an economy with money-like properties (a currency for exchange of goods and services), ver-

sus using virtual points to create psychological rewards (such as status)? Second, how do social-psychological rewards, even individual ones, aggregate in terms of the perceived value to contributors? While utility from money— both in terms of value as a function of total wealth, and the change in value of wealth with time— is a relatively well-studied subject in the economics literature, very little is known or understood about how social-psychological rewards aggregate, and how they retain (or gain or lose) value over time; also, unlike financial rewards, this could be partially controlled by system design. Understanding how multiple rewards influence incentives when they occur simultaneously in a system, and how social-psychological rewards provide value —starting with understanding agent preferences from a behavioral economics perspective, and then integrating this understanding into formal game-theoretic models— is an essential component to a strong foundation for incentive design for human computation, and one of the most exciting directions for future work in this area.

# References


[AFPT11] N. Alon, F. Fischer, A. Procaccia, and M. Tennenholtz. Sum of us: strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, 2011.

[AHKL13] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Steering user behavior with badges. In *22st International World Wide Web Conference (WWW'13)*, 2013.

[DG13] A. Dasgupta and A. Ghosh. Crowdsourced judgment elicitation with endogenous proficiency. In *Proc. 22nd ACM International World Wide Web Conference (WWW), 2013*, 2013.

[EG13] D. Easley and A. Ghosh. Incentives, gamification, and game theory: An economic approach to badge design. In *Proc. 14th ACM Conference on Electronic Commerce (EC), 2013*, 2013.

[GH12] A. Ghosh and P. Hummel. Implemeting optimal outcomes in social computing. In *Proc. 21st ACM International World Wide Web Conference (WWW), 2012*, 2012.

[GM12] A. Ghosh and R.P. McAfee. Crowdsourcing with endogenous entry. In *Proc. 21st ACM International World Wide Web Conference (WWW), 2012*, 2012.

[HZVvdS12] C. Ho, Y. Zhang, J. Vaughan, and M. van der Schaar. Towards social norm design for crowdsourcing markets. In *Proc. AAAI Workshop on Human Computation*, 2012.

[IPW10] P. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*, 2010.

[JCP12] Shaili Jain, Yiling Chen, and David Parkes. Designing Incentives for Online Question-and-Answer Forums. *Games and Economic Behavior*, 2012. Forthcoming.

[JMM12] Lian Jian and Jeffrey K. MacKie-Mason. Incentive-centered design for user-contributed content, 2012.

[JP13] S. Jain and D. Parkes. A game-theoretic analysis of the esp game. *ACM Transactions on Economics and Computation*, Jan 2013.

[KRK+12] R. Kraut, P. Resnick, S. Kiesler, Y. Ren, Y. Chen, M. Burke, N. Kittur, J. Riedl, and J. Konstan. *Building Successful Online Communities: Evidence-Based Social Design.* The MIT Press, 2012.

[MRZ05] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, pages 1359–1373, 2005.

[NRTV07] N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani. *Algorithmic Game Theory.* Cambridge University Press, New York, NY, USA, 2007.

[PPR+11] Galen Pickard, Wei Pan, Iyad Rahwan, Manuel Cebrian, Riley Crane, Anmol Madan, and Alex Pentland. Time-critical social mobilization. *Science*, 334:509–512, 2011.

[vAD08] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8), 2008.

[WRV08] I. Weber, S. Robertson, and M. Vojnovic. Rethinking the esp game. 2008. Technical Report, Microsoft Research.