Learning and Incentives in User-Generated Content: Multi-Armed Bandits with Endogenous Arms

Arpita Ghosh* Patrick Hummel†

August 3, 2012

Abstract

Motivated by the problem of learning the qualities of user-generated content on the Web, we study a multi-armed bandit problem where the number and success probabilities of the arms of the bandit are *endogenously* determined by strategic agents in response to the incentives provided by the learning algorithm. We model the contributors of user-generated content as attention-motivated agents who derive benefit when their contribution is displayed, and have a cost to quality, where a contribution's quality is the probability of its receiving a positive viewer vote. Agents strategically choose whether and what quality contribution to produce in response to the algorithm that decides how to display contributions. The algorithm, which would like to eventually only display the highest quality contributions, can only learn a contribution's quality from the viewer votes the contribution receives when displayed. The problem of inferring the relative qualities of contributions using viewer feedback, to optimize for overall viewer satisfaction over time, can then be modeled as the classic multi-armed bandit problem, *except* that the arms available to the bandit and therefore the achievable regret are endogenously determined by strategic agents — a good algorithm for this setting must not only quickly identify the best contributions, but also *incentivize* high-quality contributions to choose amongst in the first place.

We first analyze the well-known UCB algorithm $\mathcal{M}_{\text{UCB}}$ [Auer *et al.* 2002] as a *mechanism* in this setting, where the total number of potential contributors or arms, K, can grow with the total number of viewers or available periods, T, and the maximum possible success probability of an arm, γ, may be bounded away from 1 to to model malicious or error-prone viewers in the audience. We first show that while $\mathcal{M}_{\text{UCB}}$ can incentivize high-quality arms and achieve strong sublinear equilibrium regret when $K(T)$ does not grow too quickly with T, it incentivizes very low quality contributions when $K(T)$ scales proportionally with T. We then show that modifying the UCB mechanism to explore a randomly chosen restricted subset of $\sqrt{T}$ arms provides excellent incentive properties — this modified mechanism achieves strong sublinear regret, which is the regret measured against the *maximum* achievable quality γ, in *every* equilibrium, for all ranges of $K(T) \leq T$, for all possible values of the audience parameter γ.

*Cornell University, Ithaca, NY, USA, *E-mail address*: arpitaghosh@cornell.edu.

†Google Inc., Mountain View, CA, USA, *E-mail address*: phummel@google.com. This research was completed while the authors were affiliated with Yahoo! Research.

1 Introduction

User-generated content, such as reviews on sites like Yelp and Amazon, answers on online Q&A forums like Y! Answers, Quora or StackOverflow, comments on news stories or blogs, and user-created videos on YouTube or articles, for example, on Associated Content, are now ubiquitous on the web. This user-generated content (UGC) spans a vast spectrum of qualities — a small number of comments on an article (*e.g.*, on Slashdot) or answers to a question might be very thoughtful and well-written, while others range all the way from mediocre to junk. A website that hosts such user-generated content would like to sort through these contributions and selectively present only the best few contributions to each viewer to optimize her experience. However, websites typically cannot directly observe the qualities of contributions, but rather must learn their qualities from feedback from these same viewers — in most UGC settings, a viewer can provide feedback about the contributions they see, for example, by using thumbs-up style buttons. How should a website decide which contributions to display to each viewer so as to quickly learn the contributions' qualities, while still ensuring the best possible viewer experience overall?

The simplest abstraction of the problem of inferring the relative qualities of contributions, each of which can be thought of as a biased coin whose success probability is this unknown quality, while maximizing the number of successes obtained over time, is the classic multi-armed bandit problem [1]. However, unlike in the standard multi-armed bandit setting, the set of contributions, or arms available to the bandit, is not a fixed, exogenous quantity — rather, the contributions are produced by users who have a choice about *whether* to produce a contribution (for example, whether to write an article or review, or to answer a question), as well as about its *quality*, *i.e.*, how much effort to put into the contribution. Since these users would all like their contributions to be seen by a large number of viewers[1], how contributions are explored and displayed now constitutes a *mechanism* for allocating attention, which might affect the incentives of contributors to participate, as well as the quality of their contributions.

We therefore have a multi-armed bandit problem where the number and success probabilities of arms of the bandit are now *endogenously* determined in response to the incentives provided by the learning algorithm to agents. A good mechanism in this setting not only needs to quickly identify and exploit the best arms, but also must *incentivize* high quality arms, or contributions, in the first place. What algorithms provide strong incentives to users to participate and make high-quality contributions, and then also quickly identify the best content?

Our contributions. We model and analyze the problem of learning the qualities of user-generated content, when this content is produced by strategic, attention-motivated agents with a cost to quality. Our model leads to a new version of the multi-armed bandit problem with *endogenous* arms, where the number and success probabilities of the arms presented to the bandit are not fixed apriori, but rather are determined by the incentives provided by the explore-exploit algorithm.

We begin by analyzing the well-known UCB algorithm [1] $\mathcal{M}_{\text{UCB}}$ for the multi-armed bandit problem as a *mechanism*, in a setting where the maximum success probability of arm, γ, may be bounded away from 1 — this models an important factor in the context of user-generated content, namely the presence of voters in the online audience who (either deliberately or accidentally) vote down even the best contributions. We find that while the UCB mechanism is able to incentivize high-quality arms and achieve strong sublinear regret (Definition 2.2) when the number of potential contributors $K(T)$ does not grow too quickly with the number of viewers T, it fails miserably when $K(T)$ scales up with T, which is a natural regime of interest in the UGC setting where often many viewers are also potential contributors — when $\lim_{T\to\infty} \frac{T}{K(T)} = r$, *all* contributors produce very low quality content under $\mathcal{M}_{\text{UCB}}$, leading to poor regret.

We then investigate a modification of the UCB mechanism, motivated by another bandit algorithm $\mathcal{M}_{1-\text{FAIL}}$ [6] that is designed to achieve good performance (algorithmically) in the large K regime. This mechanism $\mathcal{M}_{1-\text{FAIL}}$ itself does not suffice, since while it produces strong sublinear regret for $\gamma = 1$, it is not robust to any down-voting errors by the audience, and fails to achieve strong sublinear regret for *any* $\gamma < 1$,

[1] In addition to attention (*i.e.*, number of views) being a direct psychological incentive, there are settings, such as Associated Content, where the author of an article is paid according to the number of views her article receives, so that attention might also translate to a direct monetary incentive.

when $K(T)$ grows proportionately with T. However, modifying the UCB mechanism to explore a randomly chosen restricted subset of $\sqrt{T}$ arms as in $\mathcal{M}_{1-\text{FAIL}}$, but using the UCB index, provides excellent incentive and robustness properties — this modified mechanism achieves strong sublinear regret, which is the regret measured against the *maximum* achievable quality γ, in *every* equilibrium, for all ranges of $K(T) \leq T$, for all possible values of the audience parameter γ.

Related work. There is now an enormous literature on algorithms for learning in the multi-armed bandit setting which is too large to describe properly in this paper; see, for example, [18] and references therein for a nice overview. A large part of this literature addresses settings with a fixed and finite set of arms. In the context of user-generated content with unknown quality, however, the number of contributions, or arms, is potentially related to the number of viewers, or periods available for exploration, since a page with a larger number of viewers potentially also has a larger number of contributors. Our setting is therefore most closely related to the models that do not assume there is a fixed finite number of arms, as in [6], [20], and [21]. All of these papers take a purely algorithmic approach to the learning problem, and so do not address a situation where the number and success probabilities of the arms are *endogenously* determined in response to the incentives created by the learning algorithm.

There is a small but growing literature that addresses incentives in the multi-armed bandit setting. [2], [3], [9], [10], and [13] study multi-armed bandit problems in the context of online advertising with strategic advertisers, and investigate the design of truthful mechanisms when the system must learn the (unknown) click-through rates of ads. The strategic choice of the agents in these papers are the agents' *bids*, rather than the click-through rates of the arms that are learned by the mechanism. [5] considers questions related to endogenous pricing decisions of firms when these decisions may influence whether consumers explore other firms and the consumers face a multi-armed bandit problem. [7] addresses incentives to explore in a multi-armed bandit problem when agents will learn from other agents' explorations and agents potentially have an incentive to free ride. [14] considers questions related to how agents would choose to explore when an agent's payoff depends in part on how many other agents chose to explore the same arm. Lastly, [19] and [22] consider models of worker job search and job quitting decisions in the context of a multi-armed bandit problem. However, none of these papers studies a multi-armed bandit setting in which the success probabilities of the arms themselves are endogenously chosen by strategic agents.

There is a growing body of work on incentives and strategic behavior in the context of online user-generated content [8, 11, 12, 15, 16]. [8] considers incentives for users to contribute content soon rather than delaying their contributions in the context of online Q&A sites; however, this work does not address questions related to the qualities of the contributions. [11] and [12] address the question of incentivizing high-quality user-generated content in a model very similar to that used in this paper. However, this work does not consider the incentives provided to agents in the process of learning or estimating the qualities of their contributions, nor does it attempt to use as small a number of viewers as possible to estimate the qualities of contributions before the best contribution is identified. This aspect of maximizing viewer welfare is captured by the *regret* measure which is used to quantify the performance of the mechanisms we study in this paper.

2 Model and Preliminaries

In this section, we present a model for the problem of learning the unknown qualities of user-generated contributions to achieve the best overall viewer experience, when contributors are strategic, and viewers provide feedback on the quality of the contributions that are displayed to them. The outline of the model is as follows: (i) Contributors decide on whether, and with how much effort, to write a contribution, based on how often their contribution is likely to be displayed by the algorithm (ii) Viewers vote on contributions, and the learning mechanism would like to use these votes to find the best, or top few, contributions and eventually display only these top contributions to viewers (iii) The mechanism's performance is measured by its equilibrium regret, which measures the regret based on the elicited qualities of contributions displayed in equilibrium against the best possible quality a contributor could have chosen. The model is described in detail below. (We note that this model does not attempt to capture every nuance of the UGC setting, and

hope that it will provide a building block to address other aspects of the problem of learning qualities in the UGC setting, such as sequential contributions, or the fact that a viewer may not always provide feedback on the contribution she sees; see §5 for a discussion.)

Content. We define the *quality* q of a contribution as the probability that a viewer will like the contribution, *i.e.*, rate it positively as 'good' or 'useful' or give it a thumbs-up; such ratings and thumbs-up buttons are widely used for feedback with online user-generated content. Since q is a probability, it must lie between 0 and 1. The qualities q_i of contributions are not directly observable (*i.e.*, are initially unknown) to the system, which can only infer these qualities from the viewer ratings.

There is a stream of T users, each of whom views contributions. For clarity of exposition, we begin by assuming that each user will be shown exactly one contribution (we later show how to extend all our results to displaying multiple contributions in §4.2). Each user provides feedback on the quality of the contribution he sees by giving it either a positive or negative vote.

We would like to model the fact that an audience of viewers need not be perfect: for example, it is possible that there are viewers who will always vote a contribution as bad no matter how good the contribution is, or that viewers make errors when rating contributions. A simple way to model such imperfect viewer populations is via an upper bound γ on the probability of receiving a positive vote from a random member of the audience. That is, we constrain the qualities of contributions as

$$q \in [0, \gamma],$$

where $\gamma \leq 1$ is the probability that a contribution of the highest possible quality receives a positive vote. When $\gamma = 1$, no segment of the viewer population makes errors or maliciously rates contributions negatively, so that a contribution of the highest possible quality always receives a positive vote from every viewer. A good mechanism should provide incentives that are robust to such realistic imperfections in the voting population, *i.e.*, that do not fail for audiences with $\gamma < 1$.[2]

Contributors. There is a pool of potential contributors, or agents, of size $K = K(T)$. The dependence of K on T reflects the fact that as the number of viewers for a site grows, the number of potential contributors can grow as well.

Each of the $K(T)$ potential contributors is a *strategic* agent who chooses whether to participate and the quality of her contribution (if participating) to maximize her expected payoff given the potential costs and benefits from contributing. We denote the probability that agent i decides to contribute by β_i and the quality she chooses when contributing by q_i. Since agents may decide whether or not to participate probabilistically, the number of actual contributors, which we denote by $k(T)$, is a random variable whose distribution depends on the total number of potential contributors K as well as their participation probabilities $\beta_1, \ldots, \beta_K$. Naturally, $k(T) \leq K(T)$.

If an agent chooses not to contribute, then she incurs no cost but also receives no benefit, so her net payoff is 0. An agent who contributes has a cost to quality which depends on her type, and derives benefit from attention, as described next.

The *cost* incurred by a contributor depends on the quality of the content she chooses to produce. To model the fact that different agents might have different abilities, *i.e.*, need to exert different levels of effort to produce the same quality contribution, we suppose that each agent can be one of several possible types τ in some finite set $\mathcal{T}$, each of which corresponds to a different cost function. An agent's type τ is an independent and identically distributed draw from a distribution ρ over $\mathcal{T}$ such that the probability an agent is type τ is ρ_τ.

The cost of producing content of quality q for an agent of type τ is $c_\tau(q)$, which is an increasing function of q (*i.e.*, producing higher quality content is more costly). We make the following assumptions on the functions c_τ:

[2]One could further generalize the model by assuming that there may be some viewers who make errors by rating contributions positively, so the lowest possible quality contribution is greater than 0. However, all our results continue to hold with this change to the model, since the analysis never uses the fact that the lowest possible quality contribution is 0.

1. The cost function c_τ is continuously differentiable in q for all $q < \gamma$ and all τ.

2. $c_\tau(0) > 0$.

3. $\lim_{q\to\gamma} c_\tau(q) = \infty$ for all τ.

The assumption that $c_\tau(0) > 0$ captures the fact that entry is endogenous, *i.e.*, participating and producing the contribution of even the lowest possible quality requires more effort than not participating at all. The assumption that $\lim_{q\to\gamma} c_\tau(q) = \infty$, as in [11] and [12], says that producing content of the highest possible quality, which corresponds to making *every* viewer that votes accurately happy, is nearly impossible.[3] (We note that the assumption that $c_\tau(q) \to \infty$ is not required for our main result (Theorem 4.3), showing that the modified UCB mechanism $\mathcal{M}_{\text{UCB}-\text{MOD}}$ achieves strong sublinear regret, to hold; rather, the result holds despite this assumption.)

Agents derive *benefit* from attention, *i.e.*, when their contribution is displayed to a viewer. The total benefit to an agent is the total number of viewers who see her contribution. Recall that there are a total of T such viewers, *i.e.*, T is the total number of times contributions will be displayed. Let t denote a generic time period between 1 and T. We use n_i^t to denote the number of views contributor i receives until the t-th period, so that n_i^T is the total number of views, or the total amount of attention received by contribution i.

We note here that our results will also hold when agents only derive value from positive feedback to the contribution, *i.e.*, from $q_i n_i^T$ rather than n_i^T alone — informally, this is because it is easier to incentivize agents to produce higher qualities when they value positive feedback than when they only value views.

The payoff of an agent who chooses quality q_i is the difference between the number of views she obtains, n_i^T (which can depend, in general, on the number and qualities of other contributors in addition to q_i), and her cost. Thus an agent's expected payoff from participating with quality q_i if the agent's type is τ is

$$u_i = E[n_i^T(q_i, q_{-i}, k(T))] - c_\tau(q_i).$$

Mechanism. A *mechanism* in this setting determines which contribution to display for each $t \le T$, and therefore the values of n_i^t, $t = 1, \ldots, T$, for all contributions i. The mechanism can choose which contribution to display based on the values of T, k (the number of contributions received), as well as the estimates of the qualities of these contributions from the (random) votes received by each contribution, and the number of times each contribution has been displayed so far (n_i^t). (We note again that the generalization to displaying multiple contributions, as is typically the case in real UGC applications, is presented in §4.2.)

The mechanism would like to eventually identify the highest quality contribution based on its estimates of qualities from the viewer votes, and then show this best contribution to every viewer. We next discuss how we measure performance — unlike in traditional multi-armed bandit settings, the performance of a mechanism here must be measured not only by how quickly the mechanism identifies and exploits the best contribution, but also on the qualities of the contributions it is able to *elicit* in equilibrium in the first place.

Solution Concept. In response to the incentives provided by the mechanism, agents strategically choose their contribution probabilities β_i and qualities q_i to maximize their utility. We use the solution concept of a free-entry Bayes-Nash equilibrium to determine the qualities and number of contributions (*i.e.*, arms), that will be elicited by the mechanism. Since agents' payoff functions are symmetric in the parameters of the game for all agents with the same type τ, we focus throughout on symmetric equilibria: in a symmetric equilibrium, all agents with the same type participate with the same probability and follow the same strategy of quality choices conditional on participating.

A *symmetric mixed strategy Bayes-Nash equilibrium* is a set of probabilities β_τ and distributions F_τ over qualities q, such that if agents' types are drawn according to the distribution ρ and all agents of type τ contribute with probability β_τ and choose a quality drawn from the CDF $F_\tau(q)$ conditional on contributing,

[3]While this model assumes that an agent knows his or her own cost function, since agents do not know the types of the other agents, agents face uncertainty about the cost functions of the other agents. However, the assumption that an agent knows his cost for making a contribution of quality q is not important for any of the analysis. If an agent faced uncertainty about how costly it would be for him to make a contribution of quality q, all the analysis in the paper would go through by simply replacing known costs with expected costs.

then no agent can increase her expected payoff by deviating from this strategy given the uncertainty the agent faces about the precise realization of the other agents' types. That is, no agent can profitably deviate by changing either her probability of participation or the distribution of qualities with which she is contributing.

Equilibrium Regret. The notion of regret, which is difference between the optimal cumulative reward (here, maximum possible viewer upvotes from all viewers) and the achieved cumulative reward (actual upvotes on the displayed contributions), is a natural measure of performance for our setting. We measure the performance of a mechanism that decides how to display contributions via its regret with respect to the *highest* possible qualities that could have potentially been chosen by agents:

Definition 2.1 (Strong regret)**.** *Consider a mechanism $\mathcal{M}$, and suppose $\mathcal{M}$ has a symmetric mixed-strategy equilibrium (β_τ, F_τ). Recall that $\gamma \leq 1$ is the highest possible quality that an agent may choose. The strong regret of the mechanism $\mathcal{M}$ in this equilibrium is*

$$R(T) = \gamma T - E[\sum_{t=1}^{T} q_t],$$

where q_t is the quality of the contribution that is displayed in period t, and the expectation is over the randomness in the mechanism as well as over the random choices of agents choosing qualities from distribution F_τ in this mixed-strategy equilibrium.

We will be particularly interested in the performance of the mechanisms we consider in the limit as $T \to \infty$, as is common in the literature on explore-exploit algorithms which attempt to minimize asymptotic regret (*e.g.* [1]) as well as in prior work on incentivizing user-generated content [11, 12]. The diverging attention regime is arguably the most important for the user-generated content setting: first, these are the situations where delivering high quality content matters the most from the perspective of viewer welfare. Second, the popular sites are the ones that draw the most attention-motivated contributors, as well as the ones that tend to attract contributions of varying quality. Indeed, tremendously large amounts of attention are not uncommon for popular content on the web; for instance, the most popular YouTube videos have been viewed over a hundred million times and even days-old trending videos have hundred of thousands of views, numbers that clearly belongs to the diverging attention regime. Other instances of user-generated content with diverging T include reviews of products, comments on popular articles, or answers on online Q&A sites that are viewed by many people over a long period of time. The notion of strong sublinear equilibrium regret, defined below, captures the performance of a mechanism in the diverging T regime, in all symmetric equilibria of the mechanism.

Definition 2.2 (Strong sublinear equilibrium regret)**.** *A mechanism $\mathcal{M}$ has strong sublinear equilibrium regret if $R(T) = o(T)$, i.e., $\lim_{T\to\infty} \frac{R(T)}{T} = 0$, in every symmetric equilibrium of $\mathcal{M}$.*

Since the qualities q are *endogenously* determined, a mechanism must be evaluated not only by whether it quickly learned which contribution is best, but also the contribution qualities that the mechanism is able to elicit: for a mechanism to have strong sublinear equilibrium regret, it must both be able to elicit some contributions with qualities that tend to γ, as well as quickly learn which are these contributions.

2.1 Preliminaries

Here we briefly state the necessary conditions for a set of participation probabilities and quality distributions $(\beta_\tau, F_\tau(q))$, in which all agents of type τ contribute with probability $\beta_\tau \in [0, 1]$ and chooses quality drawn from the distribution $F_\tau(q)$ upon participating, to constitute an equilibrium to a mechanism. These conditions, summarized below, can be derived using the same analysis as in [11]:

1. If $\beta_\tau = 0$, the expected benefit from participating for an agent of type τ must be no greater than the expected cost.

2. If $\beta_\tau \in (0, 1)$, the expected benefit from participating for an agent of type τ must equal the expected cost.

3. If $\beta_\tau = 1$, the expected benefit from participating for an agent of type τ must be no smaller than the expected cost.

4. For any q in the support of F_τ, no agent of type τ can obtain a strictly larger payoff by choosing some other quality $q' \in [0, 1)$ instead of q upon participating, given that the remaining agents use the strategy $(\beta_\tau, F_\tau(q))$.

The equilibrium conditions essentially say that in addition to receiving nonnegative payoffs (Conditions 1 and 3) that cannot be improved by choosing a different distribution of qualities (Condition 4), it must be the case that if agents participate probabilistically, then the payoff to participation must be zero, since otherwise agents could increase their payoff by either always or never participating (Condition 2). This zero-payoff condition is a common feature in models with free entry.

3 UCB Mechanism

In this section, we investigate the well-known UCB algorithm [1] as a *mechanism*, in which the number and success probabilities of the arms presented to the UCB algorithm are endogenously determined by strategic agents.

Let q_i^t denote the number of upvotes (or thumbs-ups) divided by the number of views that contribution i has received at time t, *i.e.*, the estimate of the quality of contribution i at time t. The UCB algorithm $\mathcal{M}_{\text{UCB}}$ proceeds as follows: it first displays all contributions once, and for each subsequent step t, it computes the value of the index $q_i^t + \sqrt{\frac{2 \ln T}{n_i^t}}$ for each contribution i, and displays the contribution for which this index is the largest.

Before analyzing the performance of $\mathcal{M}_{\text{UCB}}$, we first need to address the question of the *existence* of an equilibrium. The following result assures us that $\mathcal{M}_{\text{UCB}}$ indeed possesses an equilibrium, allowing us to analyze the regret of $\mathcal{M}_{\text{UCB}}$ with the equilibrium set of contributions (*i.e.*, the equilibrium number and qualities of contributions) as input arms.

Theorem 3.1 (Equilibrium Existence)**.** *There exists a symmetric mixed strategy equilibrium of $\mathcal{M}_{\text{UCB}}$ in which all contributors of type τ participate with probability β_τ and choose a quality drawn from the same cumulative distribution function $F_\tau(q)$ conditional on contributing.*

All results are proven in the appendix. It is worth noting that the proof of Theorem 3.1 does not depend in any way on the particulars of the mechanism $\mathcal{M}_{\text{UCB}}$. Instead, this proof simply appeals to fixed point arguments that would apply for virtually any reasonable mechanism. Since equilibrium existence would hold for virtually any mechanism with a virtually identical proof, we omit formal proofs of equilibrium existence for all other mechanisms considered in this paper, and simply note here that equilibria can be shown to exist in these settings with virtually identical arguments.

It is well known that the UCB algorithm achieves sublinear regret in a purely algorithmic sense [1]. We now proceed to analyze the behavior of $\mathcal{M}_{\text{UCB}}$ when treated as a mechanism with strategically determined arms. Recall that we assume throughout that $K(T)$, the number of potential contributors or arms, is less than or equal T, the number of viewers, or periods, available.

Our first theorem says that when the number of potential contributors $K(T)$ does not grow too quickly with T, there is no equilibrium in which agents choose qualities that remain bounded away from γ. This, together with Theorem 3.1 on the existence of equilibria, implies that when $K(T)$ grows adequately slowly, the UCB mechanism provides incentives for agents to participate with near-optimal quality in equilibrium, *i.e.*, $q \to \gamma$ in all equilibria. (Note that the number of potential participants $K(T)$ does not need to remain bounded as $T \to \infty$ — $K(T)$ can still diverge with T.)

Theorem 3.2. *Suppose $K(T)$ is such that $\lim_{T\to\infty} \frac{T}{K(T) \ln T} = \infty$. Then all agents participate in an equilibrium of $\mathcal{M}_{\text{UCB}}$ for sufficiently large T. Furthermore, for any fixed $q^* < \gamma$, the probability that an agent chooses quality $q \leq q^*$ in equilibrium goes to 0 in the limit as $T \to \infty$.*

Since every contributing agent chooses quality $q \to \gamma$ in every equilibrium as $T \to \infty$ and there are always contributing agents, it follows immediately that the UCB mechanism achieves strong sublinear regret for such $K(T)$. We summarize this in the following corollary.

Corollary 3.1. *Suppose* $\lim_{T\to\infty} \frac{T}{K(T)\ln T} = \infty$. *Then the UCB mechanism* $\mathcal{M}_{\text{UCB}}$ *achieves strong sublinear regret.*

However, as our next result shows, the UCB mechanism does not quite 'work' when the population of contributors scales proportionately with the number of viewers. In fact, not only does $\mathcal{M}_{\text{UCB}}$ fail to identify and exploit the best arm adequately quickly, it also manages to incentivize very low quality contributions — *every* agent who contributes chooses low quality contributions in *every* equilibrium of the mechanism.

We note that both regimes for the relative sizes of $K(T)$ and T are of interest in the context of user-generated content. In online Q&A sites such as Yahoo! Answers or StackOverflow, the number of users $K(T)$ who can answer a question is often significantly smaller than the number of users who consume the answer, possibly via a search engine. On the other hand, in settings like posts on discussion forums or comments on blogs where many consumers are also producers, the number of contributors may not be negligible compared to the number of viewers who consume the content, so that $K(T)/T$ is not vanishingly small.

Theorem 3.3. *Suppose that* $\lim_{T\to\infty} \frac{T}{K(T)} = r < \infty$. *Then for sufficiently large* T, *any equilibrium has the property that no agent of type* τ *makes a contribution with quality greater than* $q = c_\tau^{-1}(1 + c(0))$.

The proof (provided in the appendix) consists of two parts. (i) We first demonstrate that the incentives for *participation* provided to the agents are such that $\limsup_{T\to\infty} \frac{T}{E[k(T)]} < \infty$. (ii) We then show that if $\limsup_{T\to\infty} \frac{T}{E[k(T)]} < \infty$, the UCB algorithm is unable to identify and exploit the best arm, and gives all arms essentially the same amounts of attention, thereby providing incentives for poor *quality*. Note that here, the UCB mechanism fails in a purely algorithmic sense in addition to providing poor incentives — it fails to quickly identify and exploit the best contribution due to the large number of arms. We next address the question of designing a mechanism with good equilibrium regret.

4 Improving Equilibrium Regret

In this section, we address the problem of achieving strong sublinear regret for all $K(T)$. To do this, we borrow an idea from an explore-exploit algorithm $\mathcal{M}_{1-\text{FAIL}}$ [6, 18] that is designed specifically to achieve good regret for the infinite arms regime (in a purely algorithmic sense), and use it to develop a modification of the UCB mechanism which achieves strong sublinear equilibrium regret for all regimes of $K(T)$.

Before describing the modified UCB mechanism, we briefly describe the mechanism $\mathcal{M}_{1-\text{FAIL}}$ which motivates this modification. $\mathcal{M}_{1-\text{FAIL}}$ proceeds by first randomly selecting $\min\{\lfloor\sqrt{T}\rfloor, k(T)\}$ of the available contributions or arms and exploring each of these arms in turn, switching to a new arm as soon as the current arm receives a negative vote, unless this current arm has already received $\lfloor\sqrt{T}\rfloor$ or more positive votes. This continues until either all arms have received a negative vote, in which case the arm that received the largest number of positive votes is displayed for all remaining periods, or some arm receives $\lfloor\sqrt{T}\rfloor$ consecutive positive votes, in which case this arm is displayed for all remaining periods.

This algorithm $\mathcal{M}_{1-\text{FAIL}}$ itself does not, in fact, suffice for our setting with strategic contributors despite its excellent performance when arms are exogenously determined. While $\mathcal{M}_{1-\text{FAIL}}$ does achieve strong sublinear equilibrium regret when the audience parameter $\gamma = 1$ (*i.e.*, the maximum possible quality is $q = \gamma = 1$), this performance is not robust to any down-voting errors by the audience — $\mathcal{M}_{1-\text{FAIL}}$ fails to achieve strong sublinear equilibrium regret for *any* $\gamma < 1$ when $K(T)$ grows proportionally with T. We summarize the equilibrium regret behavior of the mechanism $\mathcal{M}_{1-\text{FAIL}}$ in the theorem below.

Theorem 4.1. *When* $\gamma = 1$, $\mathcal{M}_{1-\text{FAIL}}$ *achieves strong sublinear regret if there is some type* τ *for which* $c_\tau(q) = o(\sqrt{T})$ *when* $q = 1 - \Theta(\frac{1}{\ln T})$. *However, for any* $\gamma < 1$, *there exists some* $q^* < \gamma$ *such that the probability agents choose quality* $q > q^*$ *in equilibrium goes to zero in the limit as* T *goes to infinity when* $\lim_{T\to\infty} \frac{T}{K(T)} = r < \infty$.

Thus $\mathcal{M}_{1-\text{FAIL}}$ is extremely sensitive to the value of γ — agents do not produce qualities tending to γ in any equilibrium when $\lim_{T\to\infty} \frac{T}{K(T)} = r$ and $\gamma < 1$, so that $\mathcal{M}_{1-\text{FAIL}}$ cannot achieve strong sublinear regret in any equilibrium for *any* γ strictly less than 1. Thus, while providing an improvement over the UCB mechanism, $\mathcal{M}_{1-\text{FAIL}}$ does not perform well under all conditions when the arms are endogenously created by strategic agents responding to incentives of the mechanism. We now proceed with defining and analyzing the modified UCB mechanism.

4.1 The $\mathcal{M}_{\text{UCB}-\text{MOD}}$ Mechanism

We now analyze a modification of UCB which borrows the idea of exploring a restricted set of arms from $\mathcal{M}_{1-\text{FAIL}}$. The modified UCB mechanism $\mathcal{M}_{\text{UCB}-\text{MOD}}$ explores contributions using exactly the same index as the UCB mechanism, but differs from the UCB mechanism in that it explores from a smaller, randomly selected subset of arms when $K(T)$ is too large. This modified UCB mechanism $\mathcal{M}_{\text{UCB}-\text{MOD}}$ turns out to have extremely desirable incentive and learning properties: $\mathcal{M}_{\text{UCB}-\text{MOD}}$ achieves strong sublinear equilibrium regret for *all* values of $K(T)$ and the robustness parameter γ, with no restrictions on the cost functions c_τ.

Definition 4.1 ($\mathcal{M}_{\text{UCB}-\text{MOD}}$)**.** *The modified UCB mechanism $\mathcal{M}_{\text{UCB}-\text{MOD}}$ first randomly selects a subset of size $\min\{k(T), G(T)\}$ of the available contributions, where $G(T)$ is some integer-valued function. The mechanism then explores these selected contributions using $\mathcal{M}_{\text{UCB}}$.*

As we will see, when $G(T)$ is chosen so that $\lim_{T\to\infty} G(T) = \infty$ and $G(T) = o(\frac{T}{\ln T})$, there is a balance between exploring enough contributions and not exploring too many contributions which provides the right incentives for achieving good equilibrium regret. We begin the equilibrium regret analysis with an easy algorithmic lemma, which says that every contribution that is explored by $\mathcal{M}_{\text{UCB}-\text{MOD}}$ except for the one with the highest quality receives precisely $\Theta(\ln T)$ views.

Lemma 4.1. *Consider the modified UCB algorithm $\mathcal{M}_{\text{UCB}-\text{MOD}}$ with $G(T) \to \infty$ and $G(T) = o(\frac{T}{\ln T})$, and any $\delta > 0$. Then, any contribution with quality satisfying $q_i \leq q_{max}(T) - \delta$ receives $\Theta(\ln T)$ attention in expectation, where $q_{max}(T)$ is the quality of the highest-quality explored contribution.*

This result implies that an infinite number of agents (or all agents if $K(T)$ is bounded) will participate as $T \to \infty$ under $\mathcal{M}_{\text{UCB}-\text{MOD}}$, because if there are fewer than $G(T)$ participants, then an agent would strictly prefer contributing with any quality (and obtaining $\Omega(\ln T)$ attention) to not contributing.

The next theorem, which is the main component of the proof that $\mathcal{M}_{\text{UCB}-\text{MOD}}$ achieves strong sublinear equilibrium regret, says that the highest quality of an explored contribution in any equilibrium tends to γ as $T \to \infty$ with high probability, for appropriately chosen $G(T)$.

Theorem 4.2. *Suppose $G(T)$ is chosen so that $G(T) \to \infty$ and $G(T) = o(\frac{T}{\ln T})$. For any fixed $q^* < \gamma$, the probability that there is some agent explored by $\mathcal{M}_{\text{UCB}-\text{MOD}}$ who chooses quality $q > q^*$ goes to 1 as $T \to \infty$ in every equilibrium of $\mathcal{M}_{\text{UCB}-\text{MOD}}$.*

This result allows us to prove that the modified UCB mechanism $\mathcal{M}_{\text{UCB}-\text{MOD}}$ achieves strong sublinear equilibrium regret when $G(T)$ is as in Theorem 4.2:

Theorem 4.3. *Suppose $G(T) \to \infty$ and $G(T) = o(\frac{T}{\ln T})$. Then the modified UCB mechanism $\mathcal{M}_{\text{UCB}-\text{MOD}}$ achieves strong sublinear equilibrium regret for all values of γ and all $K(T) \leq T$.*

Note that the strong sublinear equilibrium regret property assures us that the regret of $\mathcal{M}_{\text{UCB}-\text{MOD}}$ with respect to the maximum achievable quality γ is sublinear in *every* equilibrium of $\mathcal{M}_{\text{UCB}-\text{MOD}}$, not just that there exists an equilibrium of $\mathcal{M}_{\text{UCB}-\text{MOD}}$ with this property.

The choice of $G(T)$, the maximum number of contributions explored by the modified UCB mechanism $\mathcal{M}_{\text{UCB}-\text{MOD}}$, can potentially have a significant effect on the incentives provided for both participation and quality. If $G(T)$ is large and the mechanism explores a larger number of contributions, then the expected benefit to having the highest quality contribution is correspondingly lower since $\mathcal{M}_{\text{UCB}-\text{MOD}}$ spends $\Theta(\ln T)$ time exploring each lower quality contribution. This leaves behind less attention for the highest quality

contribution, so agents will have a relatively lower incentive to choose higher qualities. In addition, since agents produce lower quality content when the mechanism explores more contributions, it is relatively less costly to participate and play equilibrium strategies when the mechanism explores more contributions, so agents are more likely to participate when the mechanism explores more contributions. Thus a larger value of $G(T)$ provides incentives for larger participation but lower quality contributions, and vice versa, so varying $G(T)$ could enable one to achieve various points on the trade-off curve between participation and quality. We leave the problem of precisely quantifying this tradeoff as an open direction for future work.

4.2 Displaying Multiple Contributions

We have shown so far that the modified UCB mechanism $\mathcal{M}_{\text{UCB-MOD}}$ achieves strong sublinear equilibrium regret when each viewer sees a single contribution. We now show that this result also extends to the more realistic setting where each viewer is shown multiple contributions, as is typically the case with user-generated content.

Suppose that when shown multiple contributions, a viewer views the first contribution with probability p_1, the second with probability p_2, and so on, where $p_1 \geq p_2 \geq \ldots$ reflects the fact that viewers typically read contributions from the top to the bottom of a webpage so that a contribution displayed near the top of a webpage is more likely to be viewed than a contribution lower down on the page. (This model of viewing behavior by users is widely used in the online advertising literature as well, where ads are modeled as having a position-dependent clickability that decreases with their position down the page.) As before, we assume that a viewer rates every contribution she views.

We extend the mechanism $\mathcal{M}_{\text{UCB-MOD}}$ in the obvious way to displaying multiple contributions. Let $m \geq 1$ be the number of contributions to be displayed to each viewer. At every time period t, $\mathcal{M}_{\text{UCB-MOD}}$ displays the m contributions with the largest values of $q_i^t + \sqrt{\frac{2 \ln T}{n_i^t}}$, displaying the contribution with the largest index most prominently (*i.e.*, in the top position), the contribution with the second-largest index next (in the second position), and so on.

We extend the notion of strong regret to mechanisms that display multiple contributions, also in the obvious way, as follows:

Definition 4.1. *The strong regret of a mechanism $\mathcal{M}$ which displays m contributions at each time t, in an equilibrium (β_τ, F_τ), is defined as*

$$R(T) = \sum_{j=1}^{m} p_j \gamma T - E[\sum_{t=1}^{T} \sum_{j=1}^{m} p_j q_{j,t}],$$

where $q_{j,t}$ is the quality of the contribution that is displayed in the j^{th} most prominent position in period t, and $\gamma \leq 1$ is the highest possible quality that a contribution may have. A mechanism $\mathcal{M}$ has strong sublinear equilibrium regret if $R(T) = o(T)$, (i.e., $\lim_{T \to \infty} \frac{R(T)}{T} = 0$) in every symmetric equilibrium of $\mathcal{M}$.

This definition of strong regret differs from the original definition only in that it takes into account the qualities of all the displayed contributions and weighs them by the frequency with which they are viewed. Strong regret again compares the qualities of the displayed contributions relative to the optimal solution where every displayed contribution at every time t has the maximum possible quality γ. As before, the expectation is over the randomness in the mechanism as well as over the random choices of agents choosing qualities from the distribution F_τ in this mixed-strategy equilibrium.

We now show that the results of Section 4.1 extend to this setting where the site owner displays $m \geq 1$ contributions at the same time to a viewer. First we note a lemma analogous to that in §4.1 which shows that all contributions except the best m contributions receive exactly $\Theta(\ln T)$ attention.

Lemma 4.2. *Consider the modified UCB algorithm $\mathcal{M}_{\text{UCB-MOD}}$ with $G(T) \to \infty$ and $G(T) = o(\frac{T}{\ln T})$, and any $\delta > 0$. Then any contribution with quality satisfying $q_i \leq q_m(T) - \delta$ receives $\Theta(\ln T)$ attention in expectation, where $q_m(T)$ is the quality of the m^{th}–best explored contribution.*

As in §4.1, to prove that $\mathcal{M}_{\text{UCB}-\text{MOD}}$ achieves strong sublinear equillibrium regret, we must first show that enough agents will make high quality contributions in equilibrium. We use Lemma 4.2 above to show that there will almost certainly be at least m explored agents who choose arbitrarily high qualities in equilibrium:

Theorem 4.4. *Suppose $G(T) \to \infty$ and $G(T) = o(\frac{T}{\ln T})$. For any fixed $q^* < \gamma$, the probability that there are at least m agents who are explored by $\mathcal{M}_{\text{UCB}-\text{MOD}}$ and that choose quality $q > q^*$ goes to 1 as $T \to \infty$ in every equilibrium of $\mathcal{M}_{\text{UCB}-\text{MOD}}$.*

As before, Theorem 4.4 allows us to prove that the modified UCB mechanism also achieves strong sublinear equilibrium regret when displaying multiple contributions:

Theorem 4.5. *Suppose $G(T) \to \infty$ and $G(T) = o(\frac{T}{\ln T})$. Then the modified UCB mechanism $\mathcal{M}_{\text{UCB}-\text{MOD}}$ achieves strong sublinear equilibrium regret for all values of γ and all $K(T) \leq T$.*

5 Discussion

In this paper, we modeled and investigated the problem of learning the qualities of user-generated content, which leads to a multi-armed bandit problem where the arms of the bandit are endogenously determined by the response of strategic agents to the incentives provided by the learning mechanism. There are a number of interesting directions for further work, consisting of both theoretical questions and mechanism design questions arising from more nuanced models; we discuss these below.

Our results show that a modification of the well-known UCB mechanism can achieve strong sublinear regret in all equilibria. What other learning algorithms can achieve this regret? We also note that we do not explicitly derive a quality-participation tradeoff as a function of $G(T)$ — this tradeoff can be particularly relevant in settings when m, the number of contributions to be displayed, is not fixed apriori but rather determined by a threshold on quality. An interesting family of open question is what classes of learning mechanisms provide strong incentives for eliciting arms with high success probabilities, and what mechanisms are optimal in terms of the quality-participation tradeoff.

Sequential Contributions. In our model, all agents simultaneously decide whether to contribute and the quality of their contribution if contributing in a Bayes-Nash equilibrium. This is a reasonable assumption, for instance, when it is too costly for an agent to look at all previous contributions and strategize about her own contribution based on this information (as in the case of, for example, popular news articles or blog posts which attract huge numbers of comments), or in settings like some Q&A forums where a number of answerers might be simultaneously working on the solution to a particular question (*e.g.*, in coding questions on StackOverflow). However, there are many other settings that are better suited to an alternative sequential model, in which different potential contributors arrive at different times and make decisions about their own contributions after viewing the existing set of contributions. The question of equilibrium contribution qualities in such a sequential model is an interesting, albeit possibly more challenging, direction for future work.

Enhancing the viewer model. Our model for multiple contributions assumes that the probability of viewing of contribution is independent of the qualities of the contributions displayed before it. However, in some settings a user may be more likely to want to continue to view contributions if they liked the first few contributions, whereas in other settings a user may be more likely to view additional contributions if their needs were not met by the previous contributions. Capturing these viewer behaviors in a model with multiple contributions and analyzing mechanisms to efficiently learn the qualities of contributions in these settings is another intriguing direction for further work.

Acknowledgments

We thank Anirban Dasgupta, Preston McAfee, and Michael Schwarz for helpful discussions.

References

[1] P. Auer, N. Cesa-Bianchi, P. Fischer, *Finite-Time Analysis of the Multiarmed Bandit Problem*, Machine Learning, 47, 235-256, 2002.

[2] M. Babaioff, R.D. Kleinberg, A. Slivkins, *Truthful Mechanisms with Implicit Payment Computation*, Proceedings of the 11th ACM Conference on Electronic Commerce (EC), 2010.

[3] M. Babaioff, Y. Sharma, A. Slivkins, *Characterizing Truthful Multi-Armed Bandit Mechanisms*, Proceedings of the 10th ACM Conference on Electronic Commerce (EC), 2009.

[4] J.G. Becker, D.S. Damianov, *On the Existence of Symmetric Mixed Strategy Equilibria*, Economics Letters, 90(1), 84-87, 2006.

[5] D. Bergemann, J. Välimäki, *Learning and Strategic Pricing*, Econometrica, 64(5), 1125-1149, 1996.

[6] D.A. Berry, R.W. Chen, A. Zame, D.C. Heath, L.A. Shepp, *Bandit Problems with Infinitely Many Arms*, Annals of Statistics, 25(5), 2103-2116, 1997.

[7] P. Bolton, C. Harris, *Strategic Experimentation*, Econometrica, 67(2), 349-374, 1999.

[8] Y. Chen, S. Jain, D. Parkes, *Designing Incentives for Online Question and Answer Forums*, Proceedings of the 10th ACM Conference on Electronic Commerce (EC), 2009.

[9] N.R. Devanur, S.M. Kakade, *The Price of Truthfulness for Pay-Per-Click Auctions*, Proceedings of the 10th ACM Conference on Electronic Commerce (EC), 2009.

[10] N. Gatti, A. Lazaric, F. Trovò, *A Truthful Learning Mechanism for Multi-Slot Sponsored Search Auctions with Externalities*, Proceedings of the 13th ACM Conference on Electronic Commerce (EC), 2012.

[11] A. Ghosh, P. Hummel, *A Game-Theoretic Analysis of Rank-Order Mechanisms for User-Generated Content*, Proceedings of the 12th ACM Conference on Electronic Commerce (EC), 2011.

[12] A. Ghosh, R.P. McAfee, *Incentivizing High-Quality User-Generated Content*, Proceedings of the 20th International World Wide Web Conference (WWW), 2011.

[13] R. Gonen, E. Pavlov, *Brief Announcement: An Incentive-Compatible Multi-Armed Bandit Mechanism*, Proceedings of the 26th ACM Conference on Principles of Distributed Computing (PODC), 2007.

[14] R. Gummadi, R. Johari, J.Y. Yu, *Mean Field Equilibria of Multi Armed Bandit Games*, Proceedings of the 13th ACM Confernce on Electronic Commerce (EC), 2012.

[15] S. Jain, D. Parkes, *The Role of Game Theory in Human Computation Systems* (Position Paper), Proceedings of the 1st Human Computation Workshop (HCOMP), 2009.

[16] S. Jain, D. Parkes, *A Game-Theoretic Analysis of the ESP Game*, ACM Transactions on Economics and Computation (TEAC), in press.

[17] S.M. Kakade, I. Lobel, H. Nazerzadeh, *An Optimal Dynamic Mechanism for Multi-Armed Bandit Processes*, Working Paper.

[18] R.D. Kleinberg, *Online Decision Problems with Large Strategy Sets*, MIT Ph.D. Thesis, 2005.

[19] B.P. McCall, J.J. McCall, *A Sequential Study of Migration and Job Search*, Journal of Labor Economics, 5(4), 452-476, 1987.

[20] O. Teytaud, S. Gelly, M. Sebag, *Anytime Many-Armed Bandits*, Conférence d'Apprentissage, 2007.

[21] Y. Wang, J. Audibert, R. Munos, *Algorithms for Infinitely Many-Armed Bandits*, Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS), 2008.

[22] W.K. Viscusi, *Job Hazards and Worker Quit Rates: An Analysis of Adaptive Worker Behavior*, International Economic Review, 20(1), 29-58, 1979.

APPENDIX

A Proofs of Results in Main Text

Proof of Theorem 3.1: First note that no player in this game would ever choose a quality $q > \max_\tau\{c_\tau^{-1}(T)\}$, as a player could always obtain a strictly greater expected payoff by not participating. Thus any mixed strategy equilibrium to the game in which players are restricted to choosing $q \in [0, \max_\tau\{c_\tau^{-1}(T)\}]$ is also a mixed strategy equilibrium of the original game.

Now note that this modified game in which players are restricted to choosing $q \in [0, \max_\tau\{c_\tau^{-1}(T)\}]$ is a symmetric game in which each player has a pure strategy space that is compact and Hausdorff. Also note that each player's expected payoff in this modified game is continuous in the actions of the players. It thus follows from Theorem 1 of [4] that there exists a symmetric mixed strategy equilibrium of this modified game. This in turn implies that there is a symmetric mixed strategy equilibrium of the original game. □

Proof of Theorem 3.2: Suppose by means of contradiction that there exists some fixed $q^* < \gamma$ and some fixed probability $\pi > 0$ such that there are infinitely many values of T for which there is an equilibrium in which agents who participate choose quality $q \leq q^*$ with probability greater than or equal to π. Throughout the remainder of the proof, restrict attention to such values of T for which the probability (unconditional on the realizations of the agent's types τ) an agent who contributes chooses quality $q \leq q^*$ is greater than or equal to π. To establish a contradiction, we will show that for sufficiently large T, an agent can obtain a strictly greater expected utility by making a contribution of quality $q^* + \epsilon$ for some small $\epsilon \in (0, \gamma - q^*)$ than by making a contribution of quality $q \leq q^*$, *i.e.*, there exists a profitable deviation.

We first show that any contributing agent obtains $\Omega(\ln T)$ attention. Since $\frac{T}{K(T)} = \omega(\ln T)$, the contribution that receives the most attention receives $\omega(\ln T)$ attention. Thus the value of $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$ for the contribution i that is explored the most tends to q_i as $t \to T$ and $T \to \infty$. Now if a contribution j is explored $o(\ln T)$ times, then the value of $q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ tends to ∞ as $T \to \infty$. This implies that if a contribution j is explored $o(\ln T)$ times, then for sufficiently large t, $q_j^t + \sqrt{\frac{2\ln T}{n_j^t}} > q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$. This contradicts the possibility that i is explored $\omega(\ln T)$ times while j is only explored $o(\ln T)$ times, so it must be the case that each contribution j is explored $\Omega(\ln T)$ times, regardless of the qualities of the contributions.

Next we show that if an agent j contributes with quality $q_j = q^* + \epsilon$ for some small $\epsilon \in (0, \gamma - q^*)$ and some other agent i contributes with quality $q_i \leq q^*$, then there must exist some $\delta > 0$ such that the probability that $\frac{n_j^t}{n_i^t} > 1 + \delta$ goes to 1 in the limit as t goes to infinity. Suppose not, so that $\lim_{t\to\infty} \frac{n_j^t}{n_i^t} \leq 1$. Then $\lim_{T\to\infty} \lim_{t\to T} \sqrt{\frac{2\ln T}{n_j^t}} - \sqrt{\frac{2\ln T}{n_i^t}} \geq 0$ for large t and $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}} < q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ for large t. But since $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}} < q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ for sufficiently large t, there exists some threshold t^* such that i is not explored for any $t \geq t^*$, which contradicts the possibility that i is explored infinitely often. Therefore, there exists some $\delta > 0$ such that the probability that $\frac{n_j^t}{n_i^t} > 1 + \delta$ goes to 1 in the limit as t goes to infinity.

But this implies that if an agent deviates and contributes with quality $q^* + \epsilon$ for some small $\epsilon \in (0, \gamma - q^*)$, then the agent can obtain $\Omega(\ln T)$ more attention in expectation than when she contributes with quality no greater than q^*. Since q^* is bounded away from γ, the cost of this deviation, which is no greater than $c(q^* + \epsilon)$, remains bounded as $T \to \infty$ although the benefit from this deviation grows with $\Omega(\ln T)$ as $T \to \infty$. Thus for sufficiently large T, an agent can profitably deviate by contributing with quality $q^* + \epsilon$ for some small $\epsilon \in (0, \gamma - q^*)$ instead of using a quality no greater than q^*. This contradicts our assumption that the postulated sequence of equilibria exists and shows that for any fixed $q^* < \gamma$, the probability the agents choose quality $q \leq q^*$ in equilibrium must go to zero in the limit as T goes to infinity.

To complete the proof, we must show that all agents participate in equilibrium for sufficiently large T. To see that this holds, recall that regardless of the qualities of the contributions, an agent's expected benefit from participating is $\Omega(\ln T)$. But if an agent contributes with quality 0, the agent pays a fixed finite cost that remains bounded as $T \to \infty$. Thus for sufficiently large T, an agent always obtains a greater expected payoff

from participating with quality 0 than not participating at all, and full participation must arise in equilibrium. □

Proof of Theorem 3.3: The proof consists of two parts. (i) We first demonstrate that the incentives for *participation* provided to the agents are such that $\limsup_{T\to\infty} \frac{T}{E[k(T)]} < \infty$ as well. (ii) We then show that if $\limsup_{T\to\infty} \frac{T}{E[k(T)]} < \infty$, the UCB algorithm is unable to identify and exploit the best arm, and gives all arms essentially the same amounts of attention, thereby providing incentives for poor *quality*.

(i) Consider a sequence of equilibria for the various values of T. Recall that the random variable $k(T)$ denotes the number of actual contributors of the $K(T)$ potential contributors, so that $E[k(T)] = \sum_\tau \beta_\tau(T)\rho_\tau K(T)$. We first show that $\limsup_{T\to\infty} \frac{T}{E[k(T)]} < \infty$ in this sequence by demonstrating that if this does not hold, a nonparticipating agent can profitably deviate by participating.

First note that if there is some subsequence of T such that there is an equilibrium for each T so that $\lim_{T\to\infty} \frac{T}{\sum_\tau \beta_\tau(T)\rho_\tau K(T)} = \infty$ holds in this subsequence, then any contributing agent must receive a diverging amount of attention in expectation in these equilibria in the limit as $T \to \infty$. To see why, note that the expected amount of attention received by the agent who receives the greatest amount of attention, say agent j, diverges in the limit as $T \to \infty$ since $\lim_{T\to\infty} \frac{T}{E[k(T)]} = \infty$. Thus if there is some contributing agent i who receives an amount of attention that remains bounded as $T \to \infty$, then $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}} > q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ for sufficiently large t and T. But this contradicts the possibility that the algorithm explores contribution j an infinite number of times in the limit as $T \to \infty$. Thus if $\lim_{T\to\infty} \frac{T}{E[k(T)]} = \infty$ along this subsequence, then any contributing agent must receive a diverging amount of attention in expectation in the limit of this sequence of equilibria.

But if any agent who participates receives a diverging amount of attention in expectation, irrespective of quality, in the limit of this sequence of equilibria, then an agent can profitably deviate by making a contribution of quality 0 and obtaining a diverging amount of attention instead of not participating at all for sufficiently large T in this subsequence. This contradiction implies that we must have $\limsup_{T\to\infty} \frac{T}{E[k(T)]} < \infty$ in any sequence of equilibria.

(ii) Next we show that if the number of arms presented to the UCB mechanism grows too quickly with T, as when $\limsup_{T\to\infty} \frac{T}{E[k(T)]} < \infty$, then for sufficiently large T the algorithm will first explore all contributions once, then explore each contribution a second time, and so on, until contributions have been displayed T times, regardless of the qualities of the contributions. To see this, note that if the algorithm displays contributions in this order and some contribution j has been displayed one more time than some other contribution i (*i.e.*. $n_j^t = n_i^t + 1$), then it is necessarily the case that $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}} > q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ for sufficiently large T. This holds because the fact that $n_j^t \leq \frac{T}{\sum_\tau \beta_\tau(T)\rho_\tau K(T)} + 1$ for large T implies there is some $s < \infty$ such that $n_j^t < s$ for all t and T, so $\sqrt{\frac{2\ln T}{n_i^t}} - \sqrt{\frac{2\ln T}{n_j^t}} \geq \sqrt{\frac{2\ln T}{s-1}} - \sqrt{\frac{2\ln T}{s}}$ for sufficiently large T. But $\sqrt{\frac{2\ln T}{s-1}} - \sqrt{\frac{2\ln T}{s}}$ becomes arbitrarily large as $T \to \infty$ since s is a constant independent of T, so $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}} > q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ for sufficiently large T. But since $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}} > q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ holds for large T whenever $n_j^t = n_i^t + 1$, if contribution j has been displayed one more time than contribution i, it is necessarily the case that contribution i must be explored at least once more before j is explored again for large T.

Thus no contribution ever obtains more than one extra unit of attention than any other contribution for sufficiently large T, irrespective of the contributions' qualities. But in this case, an agent of type τ would never choose a quality q such that $c_\tau(q) - c_\tau(0) > 1$. Thus for sufficiently large T, any equilibrium has the property that no agent of type τ makes a contribution with quality greater than $q = c_\tau^{-1}(1 + c(0))$. □

Proposition A.1. *If an agent changes her quality choice from q_i to $q_i + \epsilon$ for some small $\epsilon > 0$, then she obtains an additional success before her first failure with probability $\frac{\epsilon}{1-q_i}$.*

Proof. Let $\{r_n\}_{n=1}^\infty$ denote an infinite sequence of random variables where each r_n is an independent and identically distributed draw from the uniform distribution on $[0, 1]$. Note that if an agent makes a contribution

of quality q_i, then the probability she obtains a success on the n^{th} draw is just the probability that $r_n \leq q_i$. Similarly, the probability of obtaining a success on the n^{th} draw with a contribution of quality $q_i + \epsilon$ is just the probability that $r_n \leq q_i + \epsilon$.

Now let m denote the first draw where an agent with a contribution of quality q_i obtains a failure. Thus $r_n \leq q_i$ for all $n < m$ and $r_m > q_i$. If an agent makes a contribution with quality $q_i + \epsilon$ for some small $\epsilon > 0$, she also obtains a success for each of the first $m-1$ draws (since $r_n \leq q_i + \epsilon$ for all $n < m$), and obtains a success on the m^{th} draw if $r_m \leq q_i + \epsilon$. Since $r_m > q_i$, this agent obtains an additional success only if $r_m \in (q_i, q_i + \epsilon]$. The probability that $r_m \in (q_i, q_i + \epsilon]$ conditional on $r_m > q_i$ is $\frac{\epsilon}{1-q_i}$. Therefore, if an agent changes her quality choice from q_i to $q_i + \epsilon$ for some small $\epsilon > 0$, then she obtains an additional success before her first failure with probability $\frac{\epsilon}{1-q_i}$.

□

Proof of Theorem 4.1: First we illustrate that $\mathcal{M}_{1-\text{FAIL}}$ achieves strong sublinear regret in the case where $\gamma = 1$. To do this, let $\tilde{K}(T) = E[k(T)]$ denote the expected number of agents who participate in equilibrium. First consider the case where $\tilde{K}(T)$ remains bounded as $T \to \infty$. Note that if the 1-fail mechanism fails to achieve strong sublinear regret, there must be some $q^* < \gamma$ such that the probability all agents choose quality $q \leq q^*$ remains bounded away from zero in the limit as $T \to \infty$. In this case, if an agent i of type τ changes her quality choice from q_i to $q_i + \epsilon$ for some small $\epsilon > 0$, then she obtains an additional success before her first failure with probability $\Theta(\epsilon)$ by Proposition A.1. Furthermore, this change always has a probability bounded away from zero of affecting whether the agent obtains more successes before her first failure than any other agent (since there is a positive probability that all agents choose qualities $q_j \leq q^* < 1$). Now, if this change does affect whether the agent obtains more successes before her first failure than any other agent, then she obtains an additional $\Theta(T)$ units of attention. From this it follows that the expected benefits from making such a change are $\Theta(\epsilon T)$, so the derivative of an agent's benefit with respect to quality at q_i must be $b'(q_i) = \Theta(T)$. But the derivative of an agent's utility with respect to quality must be 0 at every quality in the support of an equilibrium distribution, *i.e.*, it is necessary that $c'_\tau(q_i) = \Theta(T)$ in equilibrium. Since $c'_\tau(q_i)$ remains bounded and finite in the limit as $T \to \infty$, this cannot hold for sufficiently large T. Thus the 1-fail mechanism achieves strong sublinear regret in this case.

Now consider the case where $\lim_{T\to\infty} \tilde{K}(T) = \infty$. Again, suppose by means of contradiction that the 1-fail mechanism fails to achieve strong sublinear regret. This implies that there is some $q^* < 1$ and some $\pi > 0$ such that $Pr(q^T_{win} \leq q^*) \geq \pi$ holds for an infinite number of T, where q^T_{win} denotes the quality that is chosen by the winning agent, *i.e.*, the agent who first has $\sqrt{T}$ consecutive successes or has the most successes before her first failure. Throughout the remainder of the proof, we restrict attention to values of T that satisfy $Pr(q^T_{win} \leq q^*) \geq \pi$.

First note that as $T \to \infty$, the probability that any agent who contributes with quality $q \leq q^*$ has $\sqrt{T}$ consecutive successes before her first failure goes to zero, so if $q^T_{win} \leq q^*$, then with probability arbitrarily close to 1, q^T_{win} is the quality chosen by the agent who has the most successes before her first failure.

Now if the mechanism explores $\min\{k(T), \lfloor\sqrt{T}\rfloor\}$ agents who all contribute with quality $q \leq q^*$, then the maximum number of consecutive successes that any such agent has before her first failure is $O(\ln T)$ with probability arbitrarily close to 1. To see this, first note that when restricting attention to $\min\{k(T), \lfloor\sqrt{T}\rfloor\}$ agents who contribute with quality $q \leq q^*$, this maximum number of successes will be largest when there are $\lfloor\sqrt{T}\rfloor$ agents and all of these agents contribute with quality $q = q^*$. Also note that if an agent contributes with quality q^*, then the probability the agent has less than r consecutive successes before her first failure is $1 - (q^*)^r$. Therefore, the probability that all $\lfloor\sqrt{T}\rfloor$ agents with quality q^* explored by the mechanism have less than r consecutive successes before their first failure is $(1 - (q^*)^r)^{\lfloor\sqrt{T}\rfloor}$. But for this probability to be bounded away from both zero and one, we must have $r = \Theta(\ln T)$. So if the mechanism explores $\min\{k(T), \lfloor\sqrt{T}\rfloor\}$ agents who contribute with quality $q \leq q^*$, then the maximum number of consecutive successes before the first failure in this group of $\min\{k(T), \lfloor\sqrt{T}\rfloor\}$ agents is $O(\ln T)$ with probability arbitrarily close to 1.

But if an agent contributes with quality q satisfying $\frac{1}{1-q} = \Theta(\ln T)$, then the expected number of successes this agent obtains before her first failure is $\Theta(\ln T)$. Therefore, if an agent contributes with quality q satisfying $\frac{1}{1-q} = \Theta(\ln T)$, then the probability this agent will have more consecutive successes before her first failure than

$\min\{k(T), \lfloor\sqrt{T}\rfloor\}$ separate agents who contribute with quality $q \leq q^*$ remains bounded away from zero in the limit as $T \to \infty$. Thus if an agent contributes with quality q satisfying $\frac{1}{1-q} = \Theta(\ln T)$, then the expected amount of attention this agent obtains if the agent is explored is $\Theta(T)$.

But since $Pr(q^T_{win} \leq q^*) \geq \pi$ holds for an infinite number of T, it must be the case that the probability any given type contributes with quality $q > q^*$ goes to zero in the limit as $T \to \infty$. Thus if τ denotes some type for which $c_\tau(q) = o(\sqrt{T})$ when $q = 1 - \Theta(\frac{1}{\ln T})$, then for sufficiently large T, there is a positive probability that some agent of type τ does not contribute with quality $q > q^*$. Also, since $\tilde{K}(T) \to \infty$, the probability such an agent obtains more successes before her first failure than any other agent conditional on being explored goes to zero in the limit as $T \to \infty$. Thus if such an agent instead contributes with a quality q satisfying $\frac{1}{1-q} = \Theta(\ln T)$, then the expected additional amount of attention this agent obtains if she is explored is $\Theta(T)$.

Now the probability an agent who contributes is explored is $\Omega(\frac{1}{\sqrt{T}})$. Combining this with the result in the previous paragraph shows that if an agent of type τ contributes with a quality q satisfying $\frac{1}{1-q} = \Theta(\ln T)$ instead of using a quality $q \leq q^*$, then the agent obtains an expected additional amount of attention $\Omega(\sqrt{T})$. But since $c_\tau(q) = o(\sqrt{T})$ when $q = 1 - \Theta(\frac{1}{\ln T})$ and $\frac{1}{1-q} = \Theta(\ln T)$, it follows that the expected additional amount of attention an agent obtains from making this change exceeds the expected costs from making this change for sufficiently large T. Thus some agent would have a profitable deviation, giving a contradiction which establishes that the 1-fail mechanism achieves strong sublinear regret when $\gamma = 1$.

Next we show that if $\gamma < 1$ and $\lim_{T\to\infty} \frac{T}{K(T)} = r < \infty$, then there exists some $q^* < \gamma$ such that the probability agents choose quality $q > q^*$ in equilibrium goes to zero in the limit as T goes to infinity. First we show that $E[k(T)] = \sum_\tau \beta_\tau(T)\rho_\tau K(T) \to \infty$ in the limit as $T \to \infty$ by arguing that if only a finite number of agents participate, then an agent who is not participating could profitably deviate by participating with quality $\frac{\gamma}{2}$. Since there are only a finite number of competing agents, if an agent participates with this quality, the agent will always have a probability bounded away from zero of having more successes before her first failure than any other agent (even if all the other agents used quality γ). Thus it cannot be the case that $E[k(T)]$ stays bounded away from infinity in the limit as $T \to \infty$.

Now suppose by means of contradiction that there is no $q^* < \gamma$ such that the probability agents choose quality $q > q^*$ in equilibrium goes to zero as T goes to infinity. In this case, note that there exists some sequence $\{\hat{q}(T)\}_{T=1}^{\infty}$ such that $\lim_{T\to\infty} \hat{q}(T) = \gamma$ and $\lim_{T\to\infty} Pr(q^T_{win} > \hat{q}(T)) = 1$, where q^T_{win} denotes the quality that is chosen by the agent who first has $\sqrt{T}$ consecutive successes or has the most successes before her first failure.

To see why, note that as $T \to \infty$, the probability that any agent has $\sqrt{T}$ consecutive successes before her first failure goes to zero (since all agents choose some quality $q \leq \gamma < 1$), so with probability arbitrarily close to 1, q^T_{win} is the quality that is chosen by the agent who has the most successes before her first failure. Let $m(T)$ denote the number of successes obtained by an agent who has the most successes before her first failure. Since the mechanism explores an arbitrarily large number of agents as $T \to \infty$ (it explores $\min\{k(T), \sqrt{T}\}$ agents), $m(T)$, the highest number of consecutive successes amongst these explored agents, becomes unbounded with probability arbitrarily close to 1. But if $m(T)$ becomes unbounded, it is exponentially more likely that an agent with some high quality q_H will achieve $m(T)$ successes before her first failure than it is that an agent with some lower quality $q_L < q_H$ achieves $m(T)$ successes before her first failure. So if there are enough agents who participate with qualities in $[\hat{q}(T), \gamma]$ (which happens by assumption and since $E[k(T)] \to \infty$), then the probability the agent who has the most successes before her first failure has quality in $[\hat{q}(T), \gamma]$ goes to 1 as $T \to \infty$. Thus such a sequence $\{\hat{q}(T)\}_{T=1}^{\infty}$ exists.

Now consider the costs and benefits to participating with quality $q \in [\hat{q}(T), \gamma]$. Recall that the number of agents who participate with quality $q \in [\hat{q}(T), \gamma]$ becomes unbounded with probability arbitrarily close to 1 as $T \to \infty$, and let $F^\tau_T(q)$ denote the cumulative distribution function corresponding to the mixed strategy equilibrium quality choices of the agents of type τ who participate in the mechanism for a given T. Conditional on an agent participating with quality $q \in [\hat{q}(T), \gamma]$, the probability this agent is explored and obtains the most successes before her first failure is $\Theta(\frac{1}{\sum_\tau \beta_\tau(T)(1-F^\tau_T(\hat{q}(T)))\rho_\tau K(T)})$ since we have seen that the agent with the most successes before her first failure will almost certainly have quality $q \in [\hat{q}(T), \gamma]$ and (unconditional on the agents' precise quality choices and types) all agents who participate with qualities $q \in [\hat{q}(T), \gamma]$ have equal chances of being this agent. Also note that if an agent is explored and obtains the most successes before

her first failure, then she obtains a benefit $\Theta(T)$. For this reason, conditional on contributing with quality $q \in [\hat{q}(T), \gamma]$, the expected benefit the agent obtains is $\Theta(\frac{T}{\sum_\tau \beta_\tau(T)(1-F_T^\tau(\hat{q}(T)))\rho_\tau K(T)})$.

Also note that we must have $\lim_{T\to\infty} \beta_\tau(T) = 0$ for all τ: if there is some τ for which $\beta_\tau(T)$ remains bounded away from zero for an infinite number of T, then the expected benefits to contributing with a quality drawn randomly according to the mixed strategy equilibrium (unconditional on the agent's type) would be $\frac{T}{\sum_\tau \beta_\tau(T)\rho_\tau K(T)} = \Theta(\frac{T}{K(T)}) = \Theta(1)$ as $T \to \infty$. But the expected cost to contributing according to the mixed strategy equilibrium (unconditional on an agent's type and quality choice) would diverge as $T \to \infty$. Thus the costs to participating would exceed the benefits for sufficiently large T, and we could not have such an equilibrium in which there is some τ for which $\beta_\tau(T)$ remains bounded away from zero for an infinite number of T.

Since $\lim_{T\to\infty} \beta_\tau(T) = 0$, we must have $\beta_\tau(T) < 1$ for all τ for sufficiently large T. But if $\beta_\tau(T) < 1$, *i.e.*, agents do not participate with probability 1, the zero payoff condition must apply, *i.e.*, the expected cost from participating equals the expected benefit for sufficiently large T. That is, we must have $E_\tau[E[c_\tau(q)|q \sim F_T^\tau(q)], q \in [\hat{q}(T), \gamma]] = \Theta(\frac{T}{\sum_\tau \beta_\tau(T)(1-F_T^\tau(\hat{q}(T)))\rho_\tau K(T)})$.

Now consider an agent who contributes with quality $q \in [\hat{q}(T), \gamma]$ and consider what happens when an agent changes her quality by ϵ for some small $\epsilon > 0$. By Proposition A.1, this change increases the number of successes the agent obtains before her first failure with probability $\Theta(\epsilon)$ if the agent is explored. Also, this change can only affect whether the agent obtains additional attention after her first failure if the agent either had $m(T)$ or $m(T) - 1$ successes. But the number of agents who obtain $m(T)$ or $m(T) - 1$ successes stays bounded in the limit as $T \to \infty$ with probability arbitrarily close to 1. Combining this with the fact that (unconditional on the precise realization of the agents' qualities and types) all agents who contribute with quality $q \in [\hat{q}(T), \gamma]$ are equally likely to obtain $m(T)$ or $m(T) - 1$ successes shows that the probability such an agent obtains $m(T)$ or $m(T) - 1$ successes is $\Theta(\frac{1}{\sum_\tau \beta_\tau(T)(1-F_T^\tau(\hat{q}(T)))\rho_\tau K(T)})$. From this it follows that if an agent changes her quality by ϵ for some small $\epsilon > 0$, the probability this affects whether she obtains additional attention after her first failure is $O(\frac{\epsilon}{\sum_\tau \beta_\tau(T)(1-F_T^\tau(\hat{q}(T)))\rho_\tau K(T)})$.

The expected additional benefit from increasing quality by ϵ is therefore $O(\frac{\epsilon T}{\sum_\tau \beta_\tau(T)(1-F_T^\tau(\hat{q}(T)))\rho_\tau K(T)})$, since if an agent wins as a result of making this change, then she obtains an additional $\Theta(T)$ attention and this happens with the probability derived above. Thus in order for agents to not be able to profitably deviate by making infinitesimal changes to their quality choices, it must be the case that $E_\tau[E[c'_\tau(q)|q \sim F_T^\tau(q)]|q \in [\hat{q}(T), \gamma]] = O(\frac{T}{\sum_\tau \beta_\tau(T)(1-F_T^\tau(\hat{q}(T)))\rho_\tau K(T)})$.

Combining this with the fact that $E_\tau[E[c_\tau(q)|q \sim F_T^\tau(q)]|q \in [\hat{q}(T), \gamma]] = \Theta(\frac{T}{\sum_\tau \beta_\tau(T)(1-F_T^\tau(\hat{q}(T)))\rho_\tau K(T)})$ shows that $\frac{E_\tau[E[c'_\tau(q)|q\sim F_T^\tau(q)]|q\in[\hat{q}(T),\gamma]]}{E_\tau[E[c_\tau(q)|q\sim F_T^\tau(q)]|q\in[\hat{q}(T),\gamma]]} = O(1)$. But for any set of distributions $F_T^\tau(q)$, it must be the case that $\lim_{T\to\infty} \frac{E_\tau[E[c'_\tau(q)|q\sim F_T^\tau(q)]|q\in[\hat{q}(T),\gamma]]}{E_\tau[E[c_\tau(q)|q\sim F_T^\tau(q)]|q\in[\hat{q}(T),\gamma]]} = \infty$ because $\lim_{q\to\gamma} \log c_\tau(q) = \infty$ for all τ, meaning $\lim_{q\to\gamma} \frac{d}{dq} \log c_\tau(q) = \lim_{q\to\gamma} \frac{c'_\tau(q)}{c_\tau(q)} = \infty$ for all τ. This gives a contradiction which shows that it must be the case that there exists some $q^* < \gamma$ such that the probability agents choose quality $q > q^*$ in equilibrium goes to zero in the limit as T goes to infinity. □

Proof of Lemma 4.1: Since no more than $G(T)$ contributions are explored for any given T, the contribution i that is displayed most often must receive at least $\frac{T}{G(T)} = \omega(\ln T)$ units of attention. Therefore, the value of $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$ for this contribution i tends to q_i as $T \to \infty$.

Now if any contribution j is explored only $o(\ln T)$ times, then the value of $q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ tends to ∞ as $T \to \infty$. For sufficiently large t and T, this would imply that $q_j^t + \sqrt{\frac{2\ln T}{n_j^t}} > q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$, which would in turn mean that there is some threshold t^* such that contribution i would never be explored for all $t \geq t^*$. Therefore, it cannot be the case that i is explored infinitely often while j is only explored $o(\ln T)$ times, so it must be the case that each contribution j is explored $\Omega(\ln T)$ times, regardless of the qualities of the contributions.

Now consider a contribution with quality $q_j \leq q_{max}(T) - \delta$. Note that it cannot be the case that this contribution receives $\omega(\ln T)$ units of attention: if this contribution receives $\omega(\ln T)$ units of attention, then

$q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ tends to q_j for large t in probability. But for a contribution i with quality $q_i = q_{max}(T)$, $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$ tends to a value that is at least as large as q_i for large t in probability. Thus since $q_j \leq q_{max}(T) - \delta$, this would imply that $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}} > q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ for large t in probability. This in turn implies that there is some threshold t^* such that j is never explored for all $t \geq t^*$ in probability, meaning contribution j cannot be explored infinitely often. This contradicts the possibility that contribution j receives $\omega(\ln T)$ units of attention in expectation. Combining this with the result in the previous paragraph shows that any contribution with quality $q_j \leq q_{max}(T) - \delta$ obtains $\Theta(\ln T)$ units of attention in expectation. □

Proof of Theorem 4.2: Suppose not. Then there is some $q^* < \gamma$ such that there are infinitely many values of T for which the probability that an agent with quality $q > q^*$ is explored is less than or equal to π for some $\pi < 1$. Throughout the remainder of the proof, we restrict attention to values of T in this subsequence.

Note that if an agent chooses quality $q_j > q^* + \epsilon$ for some $\epsilon > 0$ and this agent's contribution is explored, then the agent obtains an expected amount of attention $\Theta(T)$. To see why, consider some other contribution i with quality $q_i < q^*$. This contribution i receives $O(\ln T)$ units of attention in expectation, because if i receives $\omega(\ln T)$ units of attention, then $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$ tends to q_i for large t in probability, while $q_j^t + \sqrt{\frac{2\ln T}{n_i^t}}$ tends to a value that is greater than or equal to $q_j > q_i$. But this means there is some threshold t^* such that i is never explored for all $t \geq t^*$ in probability, meaning i cannot be explored infinitely often. This gives a contradiction which shows that i receives $O(\ln T)$ units of attention in expectation.

Therefore, if an agent chooses quality $q_j > q^* + \epsilon$ for some $\epsilon > 0$ and this agent's contribution is explored and all other contributions that are explored have quality $q_i \leq q^*$, then the contributions with quality $q_i \leq q^*$ receive only a total of $O(\min\{G(T), k(T)\}\ln T)$ units of attention in expectation, so that agent j receives $\Theta(T)$ units of attention in expectation by our assumption on $G(T)$. But the probability that all other contributions that are explored have quality $q_i \leq q^*$ is at least $1 - \pi > 0$, *i.e.*, is bounded away from 0 for all T. Thus, conditional on being explored, this agent obtains an additional $\Theta(T)$ units of attention from this deviation, so that the expected additional benefit from deviating (unconditional on being explored) is $\Theta(\frac{\min\{G(T),k(T)\}T}{k(T)})$, which becomes unbounded as $T \to \infty$ since $k(T) \leq T$ and $G(T) \to \infty$.

This implies that if there is some subsequence such that the probability there is a contribution that is explored and has quality $q_i \geq q^*$ is no greater than π for some $\pi < 1$, then an agent can profitably deviate by choosing a quality $q_j > q^* + \epsilon$. Thus for any $q^* < \gamma$, the probability there is a contribution that is explored and has quality $q_i \geq q^*$ goes to 1 in the limit as $T \to \infty$. □

Proof of Theorem 4.3: Suppose by means of contradiction that $\mathcal{M}_{\text{UCB-MOD}}$ does not achieve strong sublinear regret. Then there exists some $\delta > 0$ such that the expected number of failures in $\mathcal{M}_{\text{UCB-MOD}}$ is at least $((1-\gamma)+\delta)T$ for sufficiently large T. In order for this to take place, it must be the case that $\mathcal{M}_{\text{UCB-MOD}}$ displays contributions with quality $q_i \leq \gamma - \delta$ on $\Theta(T)$ separate occasions.

Since there are no more than $G(T)$ contributions to be explored, there must be at least one contribution with quality $q_i \leq \gamma - \delta$ that is explored at least $\Theta(\frac{T}{G(T)})$ times, *i.e.*, $\omega(\ln T)$ times. For this contribution, $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$ tends to q_i for large t in probability.

Now Theorem 4.2 guarantees that there will almost certainly be at least one explored contribution with quality arbitrarily close to γ in the limit as $T \to \infty$. That is, the highest quality contribution j explored by $\mathcal{M}_{\text{UCB-MOD}}$ has a quality q_j that tends to γ as $T \to \infty$.

For this contribution, $q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ tends to a value that is no smaller than γ for large t in probability. But then $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}} < q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ for large t in probability. This implies that there is some threshold t^* such that contribution i is never explored for all $t \geq t^*$, which contradicts the possibility that the algorithm displays contribution i infinitely often. Therefore, it cannot be the case that $\mathcal{M}_{\text{UCB-MOD}}$ displays contributions with quality $q_i \leq \gamma - \delta$ on $\Theta(T)$ separate occasions in the limit as $T \to \infty$, and so the modified UCB mechanism achieves strong sublinear regret. □

Proof of Lemma 4.2: Since no more than $G(T)$ contributions are explored for any given T, the contribution i that receives the m^{th} most attention must receive at least $\frac{p_m T}{G(T)} = \omega(\ln T)$ units of attention. Therefore,

the value of $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$ for this contribution i tends to q_i as $T \to \infty$.

Now if any contribution j receives only $o(\ln T)$ attention, then the value of $q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ tends to ∞ as $T \to \infty$. For sufficiently large t and T, this would imply that $q_j^t + \sqrt{\frac{2\ln T}{n_j^t}} > q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$, which would in turn mean that there is some threshold t^* such that contribution i would never be explored for all $t \geq t^*$. Therefore, it cannot be the case that i receives an infinite amount of attention while j only receives $o(\ln T)$ attention, so it must be the case that each contribution j receives $\Omega(\ln T)$ attention, regardless of the qualities of the contributions.

Now consider a contribution with quality $q_j \leq q_m(T) - \delta$. Note that it cannot be the case that this contribution receives $\omega(\ln T)$ units of attention: if this contribution receives $\omega(\ln T)$ units of attention, then $q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ tends to q_j for large t in probability. But for a contribution i with quality $q_i = q_m(T)$, $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$ tends to a value that is at least as large as q_i for large t in probability. Thus since $q_j \leq q_m(T) - \delta$, this would imply that $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}} > q_j^t + \sqrt{\frac{2\ln T}{n_j^t}}$ for large t in probability. This in turn implies that there is some threshold t^* such that j is never explored for all $t \geq t^*$ in probability, meaning contribution j cannot receive an infinite amount of attention. This contradicts the possibility that contribution j receives $\omega(\ln T)$ units of attention in expectation. Combining this with the result in the previous paragraph shows that any contribution with quality $q_j \leq q_m(T) - \delta$ obtains $\Theta(\ln T)$ units of attention in expectation. $\square$

Proof of Theorem 4.4: Suppose not. Then there is some $q^* < \gamma$ such that there are infinitely many values of T for which the probability that at least m agents with quality $q > q^*$ are explored is less than or equal to π for some $\pi < 1$. Throughout the remainder of the proof, we restrict attention to values of T in this subsequence.

Note that if an agent chooses quality $q_j > q^* + \epsilon$ for some $\epsilon > 0$ and this agent's contribution is explored, then the agent obtains an expected amount of attention $\Theta(T)$. To see why, consider some other contribution i with quality $q_i < q^*$. This contribution i receives $O(\ln T)$ units of attention in expectation, because if i receives $\omega(\ln T)$ units of attention, then $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$ tends to q_i for large t in probability, while $q_j^t + \sqrt{\frac{2\ln T}{n_i^t}}$ tends to a value that is greater than or equal to $q_j > q_i$. But this means there is some threshold t^* such that i is never explored for all $t \geq t^*$ in probability, meaning i cannot receive an infinite amount of attention. This gives a contradiction which shows that i receives $O(\ln T)$ units of attention in expectation.

Therefore, if an agent chooses quality $q_j > q^* + \epsilon$ for some $\epsilon > 0$ and this agent's contribution is explored and there are no more than m contributions that are explored that have quality $q_i \geq q^*$, then the contributions with quality $q_i \leq q^*$ receive only a total of $O(\min\{G(T), k(T)\}\ln T)$ units of attention in expectation, so that agent j receives $\Theta(T)$ units of attention in expectation by our assumption on $G(T)$. But the probability that there are no more than m contributions that are explored and have quality $q_i \geq q^*$ is at least $1 - \pi > 0$, *i.e.*, is bounded away from 0 for all T. Thus, conditional on being explored, this agent obtains an additional $\Theta(T)$ units of attention from this deviation, so that the expected additional benefit from deviating (unconditional on being explored) is $\Theta(\frac{\min\{G(T),k(T)\}T}{k(T)})$, which becomes unbounded as $T \to \infty$ since $k(T) \leq T$ and $G(T) \to \infty$.

This implies that if there is some subsequence such that the probability there are no more than m contributions that are explored that have quality $q_i \geq q^*$ is no greater than π for some $\pi < 1$, then an agent can profitably deviate by choosing a quality $q_j > q^* + \epsilon$. Thus for any $q^* < \gamma$, the probability there are at least m contributions that are explored that have quality $q_i \geq q^*$ goes to 1 in the limit as $T \to \infty$. $\square$

Proof of Theorem 4.5: Suppose by means of contradiction that $\mathcal{M}_{\text{UCB-MOD}}$ does not achieve strong sublinear regret. Then there exists some $\delta > 0$ such that the expected number of failures in $\mathcal{M}_{\text{UCB-MOD}}$ is at least $(\sum_{j=1}^m p_j(1-\gamma) + p_m\delta)T$ for sufficiently large T. In order for this to take place, it must be the case that $\mathcal{M}_{\text{UCB-MOD}}$ displays contributions with quality $q_i \leq \gamma - \delta$ on $\Theta(T)$ separate occasions.

Since there are no more than $G(T)$ contributions to be explored, there must be at least one contribution with quality $q_i \leq \gamma - \delta$ that is explored at least $\Theta(\frac{T}{G(T)})$ times, *i.e.*, $\omega(\ln T)$ times. For this contribution, $q_i^t + \sqrt{\frac{2\ln T}{n_i^t}}$ tends to q_i for large t in probability.

Now Theorem 4.4 guarantees that there will almost certainly be at least m explored contribution with quality arbitrarily close to γ in the limit as $T \to \infty$. That is, the m^{th} highest quality contribution j explored by $\mathcal{M}_{\text{UCB}-\text{MOD}}$ has a quality q_j that tends to γ as $T \to \infty$.

For this contribution, $q_j^t + \sqrt{\frac{2 \ln T}{n_j^t}}$ tends to a value that is no smaller than γ for large t in probability. But then $q_i^t + \sqrt{\frac{2 \ln T}{n_i^t}} < q_j^t + \sqrt{\frac{2 \ln T}{n_j^t}}$ for large t in probability. This implies that there is some threshold t^* such that contribution i is never explored for all $t \geq t^*$, which contradicts the possibility that the algorithm displays contribution i infinitely often. Therefore, it cannot be the case that $\mathcal{M}_{\text{UCB}-\text{MOD}}$ displays contributions with quality $q_i \leq \gamma - \delta$ on $\Theta(T)$ separate occasions in the limit as $T \to \infty$, and so the modified UCB mechanism achieves strong sublinear regret. $\square$