Moderated Online Communities and Quality of User-Generated Content

JIANQING CHEN, HONG XU, AND ANDREW B. WHINSTON

JIANQING CHEN is an assistant professor in information systems at the Naveen Jindal School of Management at the University of Texas at Dallas and was previously an assistant professor at the Haskayne School of Business at the University of Calgary. He received his Ph.D. from McCombs School of Business at the University of Texas at Austin in 2008. His general research interests are in electronic commerce, economics of information systems, and supply chain management. His papers have been published in academic journals, including *Decision Analysis, Decision Support Systems, Economics Letters, Information Systems Research, Journal of Management Information Systems, Journal of Marketing, Journal of Marketing Research,* and *Production and Operations Management.*

HONG XU is an assistant professor in information systems at the School of Business and Management at Hong Kong University of Science and Technology. She received her Ph.D. from McCombs School of Business at the University of Texas at Austin in 2010. Her research interests include electronic commerce, economics of information systems, and auditing. Her current research focuses on the economic issues in online reputation mechanisms and strategic communications in auditing.

ANDREW B. WHINSTON is Hugh Cullen Chair Professor in the Information, Risk, and Operation Management Department at the McCombs School of Business at the University of Texas at Austin. He is also director at the Center for Research in Electronic Commerce and editor-in-chief of *Decision Support Systems*. His recent papers have appeared in *Information Systems Research, Management Science, Marketing Science, Journal of Marketing,* and *Journal of Economic Theory.* He has published over 300 papers in the major economics and management journals and has authored 27 books. In 2005, he received the Leo Award from the Association for Information Systems for his long-term research contribution to the information systems field. In 2009, he was named the Distinguished Fellow by the INFORMS Information Systems Society in recognition of his outstanding intellectual contributions to the information systems discipline.

ABSTRACT: Online communities provide a social sphere for people to share information and knowledge. While information sharing is becoming a ubiquitous online phenomenon, how to ensure information quality or induce quality content remains a challenge because of the anonymity of commentators. This paper introduces moderation into reputation systems. We show that moderation directly affects strategic commentators' incentive to generate useful information, and moderation is generally desirable to improve information quality. We find that when being moderated with different probabilities based on their reputations, commentators might display a pattern of reputation oscillation, in which they generate useful content to build up high

Journal of Management Information Systems / Fall 2011, Vol. 28, No. 2, pp. 237–268.

0742–1222 / 2011 $9.50 + 0.00.
DOI 10.2753/MIS0742-1222280209

reputation and then exploit their reputation. As a result, the expected performance from high-reputation commentators can be inferior to that from low-reputation commentators (reverse reputation). We then investigate the optimal moderation resource allocation and conclude that the seemingly abnormal reverse reputation could arise as an optimal result. Our study underscores the importance of moderation and highlights that the frequency of moderation should be properly chosen for better performance of online communities.

KEY WORDS AND PHRASES: knowledge management, moderation, online community, reputation.

THE RISE OF SOCIAL COMPUTING AND ONLINE COMMUNITIES has ushered in a new era of content delivery, where information can be easily shared and accessed. A large number of applications have emerged that facilitate collective actions for content generation and knowledge sharing. Examples include blogs [16], social networks [14], online product reviews [7, 18], wiki applications such as Wikipedia (www.wikipedia.org) [8, 17], and online forums such as Slashdot (http://slashdot.org) [3, 10]. Because of the anonymity of Internet users, however, ensuring information quality or inducing quality content remains a challenge. This paper introduces a moderation system and examines its effect on the content quality of online communities.

Information sharing and user-generated content have become ubiquitous online phenomena. For example, Wikipedia, a free online encyclopedia, is dedicated to massive distributed collaboration by allowing visitors to add, remove, edit, and change content. In online product reviews, such as those on Amazon.com, any user can post reviews on any item, even if he or she has not bought the item on Amazon.com. Online forums such as Slashdot are another example. Slashdot, a Web site that supports discussions on user-submitted news stories and articles related to technology, is one of the most frequently visited sites on the Internet. On Slashdot, all users can express their opinions simply by posting comments under a selected topic.

As these applications have gained popularity and importance, the quality of content has become a concern. On Wikipedia, readers might be provided with content that is misleading or even incorrect. Product reviews on Amazon.com can be manipulated by sellers or book publishers to boost their products. On Slashdot, commentators might post biased or useless comments. For example, some commentators could work for companies and produce purely promotional comments; other commentators might be committed to do their best but not be competent. Wikipedia is still experimenting with different approaches to ensure the quality of content [8]. As one of its cofounders pointed out, Wikipedia, as an encyclopedia, lacks both the usual review process and the respect for expertise of most encyclopedias.[1] Wikipedia has recently introduced a policy to restrict new users from making changes unaided in certain categories, such as politics [9]. Amazon.com has introduced a voting system in which consumers can vote on whether a particular review is helpful, and the vote result affects the continued

ranking of that review. Voting mitigates the information manipulation problem, but it has its pitfalls because the voting process itself can be manipulated.

Unlike many other social networks, Slashdot has constructed a reputation system and has been recognized for its quality of content. The site uses "karma" points to measure commentators' reputations based on the quality of their past comments. Karma points or reputations are clustered into a small set of labels (e.g., terrible, bad, positive, excellent). Each comment posted by a commentator receives a score ranging from –1 to 5. The score often signals the quality of the comment and affects its readership, as a comment with a higher score typically attracts more readers. Comments' default scores differ from each other according to the commentators' reputations: commentators with a good reputation receive a high default score.

One additional step taken by Slashdot beyond the regular reputation systems is its moderation system. Once a comment is posted, it may be checked or "moderated" by selected users who can change its score and assign a label, such as "informative" or "redundant." The moderation result affects the comment's score and thus its readership, and it also changes the commentator's reputation. In a sense, moderation plays the same role here as auditing in accounting. Slashdot selects moderators randomly from among eligible users and then limits the moderator status, both in number of posts to be moderated (five) and in time (three days). This restriction ensures that no moderator can have an undue effect on the system.

Similar moderation processes have been adopted by other online communities, such as Kuro5hin (www.kuro5hin.org). In fact, the moderation process was introduced mainly to screen information. As stated by the founder of Slashdot, "the purpose of moderation is to help people organize information" [3, p. 115]; it can help users "pick up hidden gems on the sandy beach of comments" [3, p. 4]. However, it seems that the actual effect of moderation is more extensive. In particular, the refined review process could have a significant effect on commentators' incentive to generate quality content.

Introducing moderation to online communities shows promise for ensuring content quality. However, little research in information systems has been done to study the effect of moderation or the design of a moderation system. Meanwhile, it has become imperative for designers to understand the effect of different designs because modern Web-based applications in many cases directly involve input from a multiplicity of agents and agent types. In the case of user-generated content, agents who post their comments have a range of backgrounds as well as various objectives that are unknown to the designers. The challenge for the designers (of Web sites such as Slashdot, Amazon.com, and Wikipedia) is to create an environment where there is an incentive for the agents to produce reliable comments, without requiring specific knowledge of the ability or objectives of the agents, so that readers can find relevant, high-quality, and reliable information in this environment. Without an expectation by readers of a reasonably high level of reliability, the site could easily lose the attention of readers and the possibility of attracting revenue sources.

This paper is in line with studies on the design of a computing environment that produces a more reliable source of information. In particular, we examine the effect

of moderation on the performance of an online community. We consider a community consisting of both dedicated and opportunistic commentators. The former behave altruistically, whereas the latter behave strategically. Commentators have reputations in the community and post their comments of different quality. A moderation system moderates each comment, and the moderation result affects both a comment's readership and the commentator's reputation.

We start with a simple case in which the moderation system monitors comments from commentators with different reputations with the same frequency. We find that moderation has a direct effect on the opportunistic commentators' incentive to exert effort. When adequate moderation is applied, opportunistic users always exert effort regardless of their reputations, whereas if moderation is very limited, they might exert no effort at all. When the level of moderation is in the middle, opportunistic commentators might adopt a strategy mixing exertion and no exertion. We demonstrate that a reputation system that includes moderation is superior to a pure reputation system in terms of the expected performance of the community.

We also consider differentiated moderation probabilities for different reputations. We find that when the moderation system monitors low-reputation commentators more carefully, commentators may display reputation oscillation. In particular, they work hard to generate useful information for building up a high reputation in one period and then exploit it in the next. In this case, the expected performance from high-reputation commentators can be inferior to that from low-reputation ones, which again illustrates the critical impact of moderation on commentators' incentives.

Finally, we discuss the optimal moderation resource allocation, which appears to be an important issue when the moderation resource is costly. We find that when dedicated commentators play a significant role in a community and opportunistic commentators are able to generate as high-quality content as the dedicated ones, optimal moderation involves either moderating all commentators equally or moderating low-reputation commentators only. In other words, it is never optimal to monitor high-reputation commentators more closely.

Information quality has been identified as an important factor in the success of information systems [5]; the quality of content is thus a natural concern for online communities. Lopes and Galletta [11] find that in the context of intrinsically motivated online content, such as entertainment or education, the perceived quality of content and provider reputation indirectly affect consumers' willingness to pay through expected benefits. A large proportion of the existing literature focuses on reputation systems of online communities. For example, Dellarocas [4] studies the reputation mechanism in eBay-like trading environments, with a focus on how mechanism parameters (e.g., a user's feedback profile) affect sellers' effort levels and market efficiency. Ba and Pavlou [1] examine whether good reputations generate product price premiums on eBay-like trading platforms. In contrast to well-understood reputation systems, the moderation system has attracted little notice. Lampe and Resnick document some observations of the moderation practice and point out that "important challenges remain for designers of such systems" [10, p. 543]. Our paper builds up a game-theoretic

model to analyze commentators' incentives and study the effect of moderation, and it aims to address why moderation works and how it can be improved.

Several authors have looked at user-generated content from different perspectives. For example, based on a unique data set from Amazon.com, Forman et al. [7] study the relationship between reviews and sales and suggest that online community users rate reviews containing identity-descriptive information more positively and that disclosure of identity information is related to online product sales. Study of online communities has also been concerned with users' motivation for their voluntary participation in and contribution to communities. Based on their data, Wasko and Faraj [15] find several factors related to users' motivation to contribute, including the perception that it enhances their professional reputation. Fang and Neufeld [6] use the theory of legitimate peripheral participation to explain the factors that influence users' long-term participation in volunteer communities, and conclude that situated learning and identity construction behaviors are positively associated with sustained participation. Benabou and Tirole [2] develop a theory of prosocial behavior to systematically explain this motivation issue. They attribute the individuals' motivation to the intrinsic value, monetary benefit, and reputation effect derived from the participation. Our paper assumes that two different types of commentators, the dedicated type and the opportunistic type, participate in online communities for their own reasons. The former behaves like altruists who may be motivated by dominant intrinsic value or reputation effect, whereas the latter acts strategically.

Our work is also related to the studies on auditing and costly verification under other settings, such as insurance contracts and owner-management contracts. For example, in an insurance setting, agents (insured) have incentive to misreport the loss from accidents, and principals (insurers) have the right to audit agents' reports. Optimal auditing has long been discussed in the literature [13]. Our study is differentiated from such literature in that we develop a repeated game with reputation to capture the unique feature of online communities, in which moderation/auditing affects not only the agents' current period payoff but also their future payoff.

The rest of the paper is organized as follows. In the next section, we lay out our model. We analyze the equilibrium effort choice under the same moderation probabilities in the third section and under differentiated moderation probabilities in the fourth section. In the fifth section, we investigate the optimal moderation resource allocation. Some extensions and discussion are offered in the sixth section. The seventh section concludes the paper.

Model

We consider an online community in an infinite-period horizon with a large number of commentators. Commentators post comments and develop reputations based on the quality of comments. At the beginning of each period, the commentators post their comments, and the comments are moderated at some point within that period. At the end of the period, the commentators' reputations are updated based on the revealed

quality of their comments as determined by the moderation. For simplicity, we assume that comments are available to readers for the current period only.

We categorize the commentators into two different types: dedicated and opportunistic. We assume that the proportion of dedicated type is μ and thus the proportion of the opportunistic type is $(1 - \mu)$. Dedicated commentators post their opinions and behave like altruists, which would be because they derive a great deal of intrinsic value from the community and are thus dedicated to the community posting. Because of the heterogeneity in the commentators' knowledge, some comments are of high quality, whereas others are of low quality. We assume that the proportion of high-quality comments or the probability of a comment being of high quality is s, $0 < s < 1$. In contrast, opportunistic commentators behave strategically. They can exert effort ($e = 1$) at cost c or exert no effort ($e = 0$) at zero cost to generate a comment. The comment with effort is of high quality with probability k, $0 < k < 1$, whereas the comment with no effort is of low quality. Cost c can be interpreted as the time that commentators spend in properly organizing their opinions or investigating the topic under discussion.

The quality of comments is unobservable to readers ex ante but is revealed once readers go over the comments. To motivate opportunistic commentators to exert effort to generate high-quality comments and to guide readers toward those of higher quality, a moderation system is implemented in which some moderators check the quality of comments (by going over the comments) and label them as either high quality or low quality. For example, on Slashdot, comments might be moderated to be "insightful" or "informative," or to be "redundant" or "off topic." Here we use "high quality" to refer to the former category and "low quality" to the latter category. With probability α ($0 \leq \alpha < 1$), a comment is moderated at the beginning of the period; otherwise, the comment is moderated at the end of the period. The former we call early moderation or moderation, and the latter we call late moderation or feedback (as if consumers report quality feedback after consuming a product). The result of early moderation affects both the number of readers of the comment and the reputation of the corresponding commentator. Late moderation affects commentators' reputations, but it does not affect the number of readers of the comment because the quality of the comment is revealed at the end of the period.

A commentator has either a high reputation or a low reputation. We consider the commentator's reputation as high if the last comment is judged to be high quality and as low if it is deemed low quality. Such an assumption is mainly for technical simplification. Because the primary purpose of our reputation system is to examine opportunistic commentators' incentive, this simple reputation measure plays an effective sanctioning role (i.e., the threat of future punishment).

We are interested in the effect of the moderation system on opportunistic commentators' behavior. We assume that opportunistic commentators derive utility from others' reading of their comments. In particular, we assume the utility is linear in the number of readers. Note that here readers might include commentators as well as *lurkers* (i.e., the users who read comments generated by commentators but do not generate comments themselves). The high-quality comments revealed by early moderation get the maximum readership, normalized to 1, and the low-quality comments revealed by

early moderation get 0 readership. For comments with late moderation, the readership level is equal to the likelihood of their being high quality, which is also termed their "expected quality." To rule out a trivial case, we assume $c < k$; otherwise, expected maximum readership cannot compensate for the effort cost, and no opportunistic commentators exert effort.

We use subscript i, $i \in \{0, 1\}$, to indicate a commentator's reputation (with 1 representing high reputation), and we denote v_i as the commentator's expected payoff. Thus, the payoffs of opportunistic commentators at period t can be formulated as follows:

$$v_1^t = \max_{e\in\{0,1\}} \alpha ek + (1-\alpha) r_1^t + \beta ek v_1^{t+1} + \beta(1-ek) v_0^{t+1} - ce \tag{1}$$

$$v_0^t = \max_{e\in\{0,1\}} \alpha ek + (1-\alpha) r_0^t + \beta ek v_1^{t+1} + \beta(1-ek) v_0^{t+1} - ce, \tag{2}$$

where β is a discount factor and r_i^t is the expected quality of comments from commentators with reputation i.

We will focus on *steady states* in which r_i^t and v_i^t are independent of time. (They, of course, depend on the state variable—reputation i.) In other words, timing does not play a role in commentators' decisions. For this reason, we simply omit the period indicator t for our discussion and rewrite the above payoff functions as

$$v_i = \max_{e\in\{0,1\}} \left[\alpha ek + (1-\alpha) r_i\right] + \beta\left[ek v_1 + (1-ek) v_0\right] - ce, \text{ for } i \in \{0,1\}. \tag{3}$$

The term in the first square bracket represents the expected payoff from the current-period readership, and the term in the second square bracket captures the future payoff.

Notice the nature of the dynamic programming in this payoff function: the current effort choice affects not only the commentator's current stage payoff but also his or her future payoff through the realized reputation. Also, it is worth pointing out that we can treat e as a continuous variable because e can also be interpreted as the probability of exerting effort in our game-theoretic framework.[2]

Equilibrium Performance

MODERATION PROBABILITIES HAVE A CRITICAL EFFECT on opportunistic commentators' incentive to exert effort. In this section, we investigate three cases where, in equilibrium, opportunistic commentators exert effort definitely, exert no effort definitely, and exert effort with some probability, respectively.

Notice that the marginal benefit from exerting effort is the probabilistic increase in the current period payoff (αk) and the increase in discounted future payoff ($\beta k(v_1 - v_0)$). On the flip side, exerting effort incurs cost c. The balance between the marginal benefit and the marginal cost is captured by the first-order derivative of the payoff function (3),

$$\alpha k + \beta k (v_1 - v_0) - c, \tag{4}$$

which determines the commentators' equilibrium choice. If the first-order derivative is positive, meaning the marginal benefit outweighs the marginal cost, the commentator

will exert effort. Otherwise, the commentator prefers not to exert effort. It is worth noting that commentators have symmetric incentives in the sense that if it is optimal for them to exert effort when their reputation is high, they also find it optimal when their reputation is low.

Meanwhile, dedicated commentators do not behave strategically, and with probability s their comments are of high quality regardless of their current reputation. Therefore, a proportion s of dedicated commentators possess high reputation.

The Equilibrium with Effort

When the probability of early moderation (*moderation probability* hereafter) is high, opportunistic commentators have great incentive to exert effort because, otherwise, their comments would fail the early moderation and thus receive no readership. More precisely, the equilibrium with opportunistic commentators exerting effort requires high moderation probabilities such that the marginal benefit outweighs the marginal cost (i.e., $\alpha k + \beta k(v_1 - v_0) - c \geq 0$).

According to Equation (3), the opportunistic commentators' expected payoffs in equilibrium are

$$\begin{aligned} v_1 &= \alpha k + (1-\alpha) r_1 + \beta\left[k v_1 + (1-k) v_0\right] - c \\ v_0 &= \alpha k + (1-\alpha) r_0 + \beta\left[k v_1 + (1-k) v_0\right] - c. \end{aligned} \tag{5}$$

The difference between these expected payoffs, $v_1 - v_0 = (1-\alpha)(r_1 - r_0)$, plays a role in determining opportunistic commentators' incentives. Notice that the difference is a function of the moderation probability. If $\alpha = 1$, then $v_1 - v_0 = 0$, which means that the expected payoffs are the same under either reputation and this case is reduced to a trivial one. In fact, $\alpha = 1$ means each comment will be moderated and the quality will be revealed immediately, and hence the payoff is solely determined by the moderation result. For this reason, under $\alpha = 1$, reputations do not matter to either readers or commentators. To exclude this trivial case, we assume $\alpha < 1$.

Recall that the proportion of dedicated commentators with high reputations is s. Proportion k of opportunistic commentators have high reputations when they exert effort. So the size of the population in high reputations will be $\mu s + (1-\mu)k$, consisting of dedicated commentators (the first term) and opportunistic commentators (the second term). Since the expected qualities of comments from dedicated commentators and from opportunistic commentators are s and k, respectively, we can formulate the expected quality of comments from high-reputation commentators as follows:

$$r_1 = \frac{\mu s s + (1-\mu) k k}{\mu s + (1-\mu) k}. \tag{6}$$

Similarly, we can formulate the expected quality of comments from low-reputation commentators as

$$r_0 = \frac{\mu (1-s) s + (1-\mu)(1-k) k}{\mu (1-s) + (1-\mu)(1-k)}.$$

Based on the expected payoff functions (5), we can rearrange the first-order incentive condition as

$$\alpha k\left[1-\beta\left(r_1-r_0\right)\right]+\beta k\left(r_1-r_0\right)-c\geq 0. \tag{7}$$

Clearly, the left-hand side is increasing in α: the higher the moderation probability, the more likely the opportunistic commentators are to exert effort. Intuitively, increasing moderation probability means increasing the chance of receiving early moderation, which encourages opportunistic commentators to exert effort because they would get caught easily and their comments would be revealed as low quality. Therefore, a higher moderation probability is more likely than a lower one to induce opportunistic commentators to exert effort.

We define α_H as the value of α that binds the above inequality (7), which is

$$\alpha_H=\frac{c/k-\beta\left(r_1-r_0\right)}{1-\beta\left(r_1-r_0\right)}=\frac{c/k-\beta\dfrac{\mu\left(1-\mu\right)\left(k-s\right)^2}{\left[k-\mu\left(k-s\right)\right]\left[1-k+\mu\left(k-s\right)\right]}}{1-\beta\dfrac{\mu\left(1-\mu\right)\left(k-s\right)^2}{\left[k-\mu\left(k-s\right)\right]\left[1-k+\mu\left(k-s\right)\right]}}. \tag{8}$$

Thus, we obtain the following lemma:

Lemma 1: Under any $\alpha\geq\alpha_H$, exerting effort can be sustained as an equilibrium.

It is worth noting that when the effort cost c is high enough (e.g., $c>k$), no moderation scheme can induce opportunistic commentators to exert effort. Recall that the maximum readership/benefit that commentators can achieve is 1 at each period. Therefore, when the cost is beyond the expected maximum readership (k), no opportunistic commentators will exert effort in any cases. This justifies our earlier assumption that $c<k$.

From the definition of α_H, $\alpha>c/k$ is a sufficient condition to induce opportunistic commentators to exert high effort. Intuitively, $\alpha>c/k$ means that the expected increase in the current period payoff (αk) outweighs the marginal cost (c), which provides commentators with adequate incentive to exert effort.

The Equilibrium with No Effort

Because dedicated commentators can have either high reputations or low reputations, readers have a certain quality expectation of the comments even from the low-reputation commentators. As a result, opportunistic commentators may catch a "free ride" on those dedicated commentators by receiving some readership without exerting any effort, as long as they are not caught in early moderation. Thus, when the moderation probability is low enough, the "free-ride" strategy would be the opportunistic commentators' best choice. More precisely, when the marginal benefit from exerting effort is not enough to compensate for the marginal cost (i.e., $\alpha k+\beta k(v_1-v_0)-c<0$), opportunistic commentators exert no effort in equilibrium. The equilibrium expected payoffs are $v_0=(1-\alpha)r_0+\beta v_0$ and $v_1=(1-\alpha)r_1+\beta v_0$. Their difference is $v_1-v_0=(1-\alpha)(r_1-r_0)$.

In this case, opportunistic commentators maintain low reputations because they exert no effort. As a result, the high-reputation commentators are composed purely of dedicated commentators, and therefore the expected quality of comment from them is s (i.e., $r_1 = s$). Low-reputation commentators consist of both dedicated commentators and opportunistic ones. Notice that a proportion $1 - s$ of dedicated commentators is in the low-reputation category with the opportunistic ones. We can then formulate the expected quality of comments from low-reputation commentators as

$$r_0 = \frac{\mu(1-s)s}{\mu(1-s)+1-\mu}. \tag{9}$$

Substituting r_1 and r_0 in the first-order derivative and rearranging the terms, we have

$$\alpha k\left[1-\beta(s-r_0)\right]+\beta k(s-r_0)-c \le 0. \tag{10}$$

Again, the left-hand side is increasing in α: the lower the moderation probability, the less likely opportunistic commentators are to exert effort. The intuition is similar to the earlier case: decreasing the moderation probability also decreases the marginal benefit from exerting effort. We define α_L as the value of α binding in the above inequality, which is

$$\alpha_L = \frac{c/k-\beta(s-r_0)}{1-\beta(s-r_0)} = \frac{c/k-\beta\dfrac{s(1-\mu)}{\mu(1-s)+1-\mu}}{1-\beta\dfrac{s(1-\mu)}{\mu(1-s)+1-\mu}}. \tag{11}$$

Thus, we can derive the following lemma:

Lemma 2: Under any $\alpha \le \alpha_L$, exerting no effort can be sustained as an equilibrium.

The intuition is as we described at the beginning of this subsection. Opportunistic commentators can expect a certain level of readership even if they do not exert any effort, as long as they do not get caught by early moderation. In this case, the level of expectation in the performance is attributed to the dedicated commentators because they always contribute. This expectation provides opportunistic commentators with a chance to free ride. When the moderation probability is low and there is only a low chance of getting caught and ending up with zero readership, opportunistic commentators have an incentive to free ride on the dedicated commentators. Thus, low moderation results in no effort.

However, when the cost of effort is low enough (such that $c/k \le \beta(s - r_0)$ and then $\alpha_L \le 0$), exerting no effort cannot be sustained as an equilibrium, no matter how low the moderation probability is. The reason is that when free riding is expected, the expected readership is also adjusted to a lower level in equilibrium. Meanwhile, opportunistic commentators always have the option of exerting effort, joining the high-reputation group, and obtaining high expected readership. When the effort cost is very low, the benefit from free riding will be exceeded by the net benefit from exerting effort. As a

result, regardless of how low the moderation probability is, opportunistic commentators choose to exert effort.

Mixed-Strategy Equilibrium

The above analysis characterizes the opportunistic commentators' equilibrium effort choice when the moderation probability is very high or very low. What will their equilibrium choice be if the moderation probability is between the two, say $\alpha_L < \alpha < \alpha_H$?[3] In such cases, we can speculate that in equilibrium some opportunistic commentators might exert effort, whereas others do not, or they sometimes exert effort but other times do not. This speculation involves mixed-strategy equilibria.

For a mixed strategy (between exerting effort and not exerting effort) to arise in equilibrium, opportunistic commentators must be indifferent about exerting effort or not; otherwise, they could always go with the more profitable option. So the marginal benefit balances the marginal cost in equilibrium; that is, $\alpha k + \beta k(v_1 - v_0) - c = 0$. We consider a symmetric case where opportunistic commentators exert effort with probability m in each reputation.[4] In such a case, the difference in expected payoffs associated with high and low reputations is again equal to the difference in the current period payoff; that is, $v_1 - v_0 = (1 - \alpha)(r_1 - r_0)$ (refer to Equation (3)). The proportion of opportunistic commentators with high reputations will be mk. Then, we can characterize the expected qualities of comments from high reputation and low reputation, respectively, as

$$r_1 = \frac{\mu ss + (1-\mu)mkmk}{\mu s + (1-\mu)mk} \tag{12}$$

$$r_0 = \frac{\mu(1-s)s + (1-\mu)(1-mk)mk}{\mu(1-s) + (1-\mu)(1-mk)}. \tag{13}$$

Based on the first-order condition, we derive the mapping between the moderation probability and the mixed strategy:

$$\alpha(m) = \frac{c/k - \beta(r_1 - r_0)}{1 - \beta(r_1 - r_0)} = \frac{c/k - \beta \dfrac{\mu(1-\mu)(mk-s)^2}{\left[mk - \mu(mk-s)\right]\left[1 - mk + \mu(mk-s)\right]}}{1 - \beta \dfrac{\mu(1-\mu)(mk-s)^2}{\left[mk - \mu(mk-s)\right]\left[1 - mk + \mu(mk-s)\right]}}. \tag{14}$$

Lemma 3: For any $\alpha \in [\alpha_L, \alpha_H]$, exerting effort with probability m can be sustained as an equilibrium, where m is determined by Equation (14).

A mixed strategy may arise as an equilibrium because of the externality of the benefit from free riding. Opportunistic commentators benefit from pooling with or free riding on dedicated commentators when they do not exert effort and do not get moderated. However, as the number of free riders increases, the readers' expectation of the pool decreases. As a result, opportunistic commentators get less readership

and less benefit from free riding. If the benefit from free riding is greater than the net benefit from exerting effort, the number of free riders will increase and thus the benefit declines. Otherwise, the number of free riders decreases and the benefit from free riding increases. In equilibrium, the benefit from free riding balances the net benefit from exerting effort, which also determines the number of free riders (or the probability that opportunistic commentators will exert effort).

In summary, we characterize the full equilibrium under different moderation probabilities in the following proposition.

Proposition 1 (Equilibrium Effort): The following describes an equilibrium: for $\alpha > \alpha_H$*, opportunistic commentators exert effort; for* $\alpha < \alpha_L$*, opportunistic commentators exert no effort; for* $\alpha \in [\alpha_L, \alpha_H]$*, opportunistic commentators exert effort with probability* $m(\alpha)$ *(determined by Equation (14)).*

Because different moderation arrangements provide different incentives for opportunistic commentators to exert effort, moderation plays a critical role in determining the equilibrium expected performance. When the moderation probabilities are the same for high and low reputations, as we have discussed so far, the equilibrium expected performances associated with each reputation appear in a uniform rank as summarized in the following proposition. This uniformity is in contrast to the case with differentiated moderation probabilities, shown in the next section.

Proposition 2 (Expected Performance): In the equilibrium as defined by Proposition 1, the expected performance of high-reputation commentators is (weakly) higher than that of low-reputation commentators. Formally, $r_1 \geq r_0$.

Proof: All proofs are presented in the Appendix, unless indicated otherwise.

This result looks very natural in that a high reputation is normally perceived as an indicator of good performance. However, it is not trivial. In our case, the expected performance of a reputation is essentially determined by the population composition (dedicated or opportunistic) under that reputation and the opportunistic commentators' performance. (Recall that dedicated commentators perform at the same level under each reputation.) Note that in each case of the equilibrium as defined by Proposition 1, opportunistic commentators exert the same level of effort under each reputation because of the symmetric incentive (which is due to the same moderation probability). Therefore, Proposition 2, in fact, says that the higher-performance commentators dominate in the high-reputation group more than in the low-reputation group.

Reputation Without Moderation

Reputation systems are used ubiquitously in online marketplaces and communities to provide information on users' abilities and trustworthiness. In most cases, however, they are not combined with a moderation system. In this subsection, we compare the moderated reputation system with a pure reputation system. Setting $\alpha = 0$ reduces the moderation system described into a pure reputation system.

Without moderation, the marginal benefit of exerting effort is from the increase in discounted future payoff ($\beta k(v_1 - v_0)$) only, which is in contrast to the increase in both the current period payoff and discounted future payoff in the case with moderation. The marginal cost is c, as before, so compared to the case with moderation, the marginal benefit from exerting effort diminishes while the marginal cost stays the same. As a result, we have:

> *Corollary 1: The overall performance under a moderation system ($\alpha > 0$) is (weakly) better than that under a pure reputation system ($\alpha = 0$).*

The corollary indicates that moderation is generally desirable for better performance in an online community, if the cost of moderation is zero or minimal. When the moderation incurs considerable cost, the extent of moderation needs to balance the cost and the benefit. Slashdot, for instance, employs a massively distributed moderation approach, in which all eligible readers have the potential to be invited as moderators, voluntarily checking or auditing for the Slashdot community. Such an arrangement provides a cost-effective way to implement a moderation system in online communities.

Differentiated Moderation Probabilities

So far, we have taken for granted that the same moderation probability is applied to commentators in both the high- and low-reputation categories. It is plausible that the community may arrange different moderation schemes for each reputation group since, after all, reputation to some degree implies commentators' types or effort. For example, the moderation system might watch low-reputation commentators more carefully, considering that they perform to a lower standard.

In this section, we study a more general case in which the moderation system moderates comments from commentators who have different reputations with different probabilities. We denote α_1 (α_0) as the moderation probability for high-(low-)reputation commentators. Replacing the moderation probability α with the differentiated probabilities α_i in the payoff function (3), we can get a similar payoff function.

The basic trade-off in commentators' decisions remains the same, except that now we have differentiated moderation probabilities. As in Equation (4), the incentive to exert effort is determined by $\alpha_i k + \beta k(v_1 - v_0) - c, i \in \{0,1\}$. Because of the differentiated moderation probabilities, unlike the previous case, opportunistic commentators might choose asymmetric effort in equilibrium: they might choose to exert effort when they are in one reputation category and choose not to do so when they are in the other reputation category.

We first consider the case $\alpha_1 < \alpha_0$, meaning the system watches low-reputation commentators more closely. Similar to the case in which there is no discrimination in moderation, we still can derive the upper bound and lower bound of the moderation probability to identify when opportunistic commentators do and do not exert effort. Note that in the current case, opportunistic commentators have asymmetric incentive to exert effort when they have different reputations. In particular, low-reputation

opportunistic commentators have more incentive to exert effort because they are more likely to get early moderation.

We are more interested in the case where opportunistic commentators adopt different strategies under different reputations. In general, more moderation gives commentators more incentive to exert effort. Given the moderation probabilities $\alpha_1 < \alpha_0$, it may arise as an equilibrium that opportunistic commentators exert no effort when possessing high reputations, whereas (some) opportunistic commentators exert effort when possessing low reputations. We first consider a steady-state equilibrium in which a relatively small proportion w ($w < k/(1 + k)$) of opportunistic commentators has high reputations and a proportion $1 - w$ has low reputations (and the number of opportunistic commentators with each reputation is invariant over time). Under such a scenario, it must be the case that low-reputation opportunistic commentators exert effort with probability $w/((1 - w)k)$ to make the number of opportunistic commentators with high reputation stable. The expected performance can be formulated as

$$r_1(w) = \frac{\mu ss}{\mu s + (1-\mu)w} \tag{15}$$

$$r_0(w) = \frac{\mu(1-s)s + (1-\mu)w}{\mu(1-s) + (1-\mu)(1-w)}. \tag{16}$$

When a relatively large proportion ($w > k/(1 + k)$) of opportunistic commentators has a high reputation in a steady-state equilibrium, it must be the case that some high-reputation and all low-reputation opportunistic commentators exert effort. If the probability of high-reputation opportunistic commentators exerting effort is x, from the steady-state condition, we have $w = (1 - w)k + wxk$ and thus $x = 1 - (k - w)/(wk)$. We can then similarly formulate

$$r_1(w) = \frac{\mu ss + (1-\mu)\left[w - (1-w)k\right]}{\mu s + (1-\mu)w} \tag{17}$$

$$r_0(w) = \frac{\mu(1-s)s + (1-\mu)(1-w)k}{\mu(1-s) + (1-\mu)(1-w)}. \tag{18}$$

We denote

$$H_1(\alpha_0 | w) \equiv \frac{\left[1 + \beta r_0(w)\right]\alpha_0 + \beta\left[r_1(w) - r_0(w)\right] - c/k}{\beta r_1(w)}, \text{ for } w \in \left[0, \frac{k}{1+k}\right]$$

and

$$H_2(\alpha_0 | w) \equiv \frac{\beta\left[k - r_0(w)\right]\alpha_0 + \beta\left[r_0(w) - r_1(w)\right] + c/k}{1 + \beta k - \beta r_1(w)}, \text{ for } w \in \left[\frac{k}{1+k}, k\right].$$

As we shall see, $\alpha_1 = H(\alpha_0 | w)$ defines a line on which w of opportunistic commentators having high reputation are sustained as an equilibrium. Such an equilibrium is characterized by the following proposition:

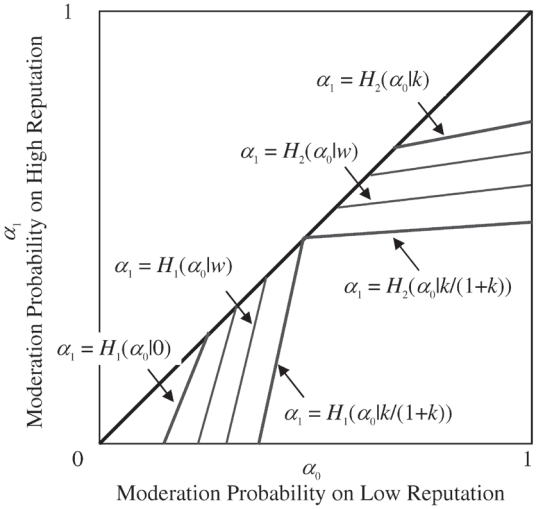

Figure 1. Moderation Supporting w of the Opportunistic in High Reputations Under $\alpha_1 < \alpha_0$

Proposition 3 (Reputation Oscillation): For any (α_0, α_1) *with* $\alpha_1 \leq \alpha_0$*, the following is an equilibrium (refer to Figure 1):*[5]

(a) If $H_1(\alpha_0|k/(1+k)) < \alpha_1 \leq H_1(\alpha_0|0)$*, a proportion* w $(0 \leq w < k/(1+k))$ *of opportunistic commentators are in high reputation and exert no effort in each period, and the other opportunistic commentators are in low reputation and exert effort with probability* $w/((1-w)k)$*, where* w *is determined by* $H_1(\alpha_0|w) = \alpha_1$*.*

(b) If $H_2(\alpha_0|k/(1+k)) < \alpha_1 \leq H_2(\alpha_0|k)$*, a proportion* w $(k/(1+k) < w \leq k)$ *of opportunistic commentators are in high reputation and exert effort with probability* $1-(k-w)/(wk)$ *in each period, and the other opportunistic commentators are in low reputation and exert effort, where* w *is determined by* $H_2(\alpha_0|w) = \alpha_1$*.*

(c) If $\alpha_1 \leq H_1(\alpha_0|k/(1+k)$ *and* $\alpha_1 \leq H_2(\alpha_0|k/(1+k)$*, a proportion* $k/(1+k)$ *of opportunistic commentators are in high reputation and exert no effort in each period, and the other opportunistic commentators are in low reputation and exert effort.*

The proposition predicts that the reputations of opportunistic commentators oscillate between high and low: they build up high reputations when they are in low-reputation states and then exploit the reputation (with some probability in Proposition 3b) when they are in high-reputation states.

As shown in Figure 1, $\alpha_1 = H_1(\alpha_0|w)$ or $\alpha_1 = H_2(\alpha_0|w)$ defines a line such that each pair of (α_0, α_1) on this line can support reputation oscillation with w of opportunistic commentators in high reputation in equilibrium.

The condition $\alpha_1 \leq H_1(\alpha_0|0)$ is to make sure that it is at least in some opportunistic commentators' interest to exert effort. In fact, when $\alpha_1 > H_1(\alpha_0|0)$, all opportunistic commentators maintain a low reputation and exert no effort. (See the bottom left-hand

corner in Figure 1.) So, similar to α_L in the case with uniform moderation probabilities, $\alpha_1 = H_1(\alpha_0|0)$ defines the boundary condition beyond which no opportunistic commentators exert effort. Similarly, $\alpha_1 = H_2(\alpha_0|k)$ defines the boundary condition beyond which all opportunistic commentators exert effort.

The above discussion shows the importance of moderation. In general, moderation plays a role in inducing opportunistic commentators' effort, and the frequency of moderation affects opportunistic commentators' incentives to exert effort. As shown, when low-reputation commentators are moderated more frequently, opportunistic commentators could optimally choose to exert more effort when they have low reputations than when they have high reputations. As a result, the overall performance of low-reputation commentators may be even better than that of high-reputation commentators. The following result exemplifies the conditions for such a circumstance:

> *Corollary 2 (Reverse Reputation): When the equilibrium w ($w \le k/(1 + k)$), determined by $H_1(\alpha_0|w) = \alpha_1$ in Proposition 3, is greater than $s\mu^{1/2}/(1 + \mu^{1/2})$, the expected performance of high-reputation commentators is lower than that of low-reputation commentators; that is, $r_0(w) > r_1(w)$.*

In these scenarios, high reputation, in fact, means something "bad" (and in equilibrium, readers anticipate that). This scenario is in sharp contrast to the standard reputation measure, where high reputation is believed to be an indicator of high quality (in adverse selection settings) or high effort (in moral hazard settings). Reputation under this moderation would be simply a symbol with no definite meaning, which again highlights the significant effect of moderation on online communities.

In a distributed moderation system, as in Slashdot, moderators may have different preferences for checking high-reputation or low-reputation comments more frequently, as there is no direct control on their preference. As a result, the moderators, overall, might check the low-reputation commentators more often. In such instances, readers should be informed of such a fact or be guided to read comments from low-reputation commentators first as reputation is a misleading indicator of comment quality.

Along a similar line, we can derive equilibria under moderation schemes with $\alpha_1 > \alpha_0$. In these cases, high-reputation commentators have more incentive to exert effort. In a steady-state equilibrium with proportion w of the opportunistic commentators in high reputation, it must be that high-reputation commentators exert effort and low-reputation commentators exert effort with probability x such that $w = wk + (1 - w)xk$ and thus $x = w(1 - k)/((1 - w)k)$. Then,

$$r_1(w) = \frac{\mu s s + (1-\mu) w k}{\mu s + (1-\mu) w} \tag{19}$$

$$r_0(w) = \frac{\mu(1-s)s + (1-\mu) w (1-k)}{\mu(1-s) + (1-\mu)(1-w)}. \tag{20}$$

The incentive conditions require that high-reputation opportunistic commentators are induced to exert effort while their low-reputation counterparts are indifferent; formally,

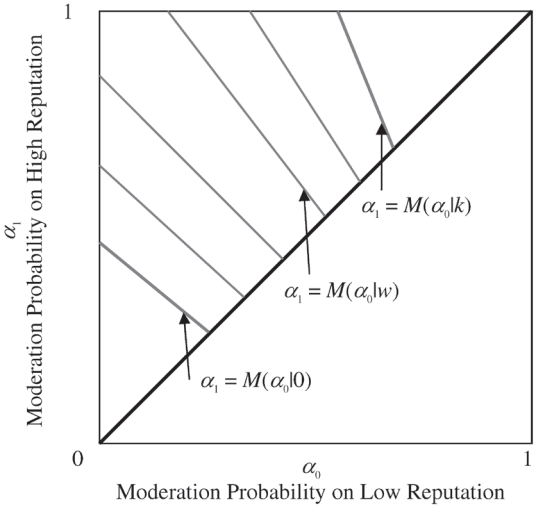

Figure 2. Moderation Supporting w of the Opportunistic in High Reputations Under $\alpha_1 > \alpha_0$

$$\begin{aligned} \beta k\left(v_1 - v_0\right) + \alpha_1 k - c > 0 \\ \beta k\left(v_1 - v_0\right) + \alpha_0 k - c = 0. \end{aligned} \tag{21}$$

Note that $v_1 - v_0 = (\alpha_1 - \alpha_0)k + (1 - \alpha_1)r_1 - (1 - \alpha_0)r_0$. By substituting in $r_i(w)$, the incentive condition for low-reputation commentators can be reorganized as

$$\alpha_1 = M\left(\alpha|w\right) \equiv \frac{-\left(1-\beta k+\beta r_0\left(w\right)\right)\alpha_0 + c/k - \beta\left(r_1\left(w\right) - r_0\left(w\right)\right)}{\beta\left(k - r_1\left(w\right)\right)}. \tag{22}$$

Similarly, $\alpha_1 = M(\alpha_0|w)$ defines a line such that each pair of (α_0, α_1) on this line can support an equilibrium with w of opportunistic commentators in high reputations (see Figure 2). When the moderation probability for high-reputation commentators is below a lower bound ($M(\alpha_0|0)$), no opportunistic commentators exert effort; and when the moderation is above an upper bound ($M(\alpha_0|k)$), all opportunistic commentators exert effort.

Optimal Moderation Allocation

When the community has adequate resources for moderation, it is always desirable to moderate the comments as much as possible. For example, if the community has a total moderation resource greater than the minimum moderation requirement needed to induce the highest effort (α_H, defined by Equation (8)), moderating comments with equal probability regardless of the commentators' reputations can induce opportunistic commentators to exert effort.

In reality, however, resources for moderation are often limited and scarce, and moderation is costly. So the community designer needs to balance the benefit of

increased overall "system performance" and the cost of attaining it. In other words, the community designer faces a decision on optimal moderation. We define the overall system performance as the expected quality of all comments, or the average quality of comments from each reputation weighted by its respective population size, $n_1 r_1 + n_0 r_0$, where n_i is the proportion of commentators with reputation i. Such a definition does measure the overall system performance because it reflects the total size of the readership of a community. Also, we assume the moderation cost is an increasing convex function of total moderation resources ($n_1\alpha_1 + n_0\alpha_0$) and denote it as $C(n_1\alpha_1 + n_0\alpha_0)$. Then, the community designer's objective function can be formulated as

$$\max_{\alpha_1,\alpha_0} \left(n_1 r_1 + n_0 r_0\right) - C\left(n_1\alpha_1 + n_0\alpha_0\right). \tag{23}$$

Note that the system performance ($n_1 r_1 + n_0 r_0$) is determined by the expected number of opportunistic commentators who exert effort in equilibrium, as dedicated commentators' performance is not affected by the moderation system design. If proportion w of opportunistic commentators maintain a high reputation over time, there must be w/k of opportunistic commentators who exert effort (considering the expected quality of a comment with effort k), and thus $n_1 r_1 + n_0 r_0 = \mu s + (1 - \mu)w$. We next examine the minimum moderation resource required to achieve a proportion w, or

$$\min_{\alpha_1,\alpha_0} \left(n_1\alpha_1 + n_0\alpha_0\right), \tag{24}$$

subject to

$$n_1 r_1 + n_0 r_0 = \mu s + \left(1 - \mu\right) w. \tag{25}$$

We are interested in whether the moderation system should moderate high-reputation commentators more or low-reputation ones more. To avoid the technical complexity, we next consider a case in which the performance of opportunistic commentators with effort is at least as good as that of dedicated ones (i.e., $k \geq s$).

Proposition 4 (Optimal Moderation): Considering the steady-state equilibrium under $k \geq s$,

(a) For all (α_0, α_1) with $\alpha_1 \leq \alpha_0$, moderating low-reputation commentators only (i.e., $\alpha_1 = 0$) is superior to any other scheme if $w \leq k/(1 + k)$; otherwise, equally moderating all commentators (i.e., $\alpha_0 = \alpha_1$) is superior to any other schemes.

(b) For all (α_0, α_1) with $\alpha_1 \geq \alpha_0$, we define w^* as the solution to

$$k - \left(\frac{1}{\beta} - 1\right) n_1\left(w\right) - \left[r_0\left(w\right) + r_1\left(w\right)\right] = 0,$$

where $n_1 = \mu s + (1 - \mu)w$, and r_1 and r_0 are defined in Equations (19) and (20), respectively. If $w > w^*$, equally moderating all commentators (i.e., $\alpha_0 = \alpha_1$) is superior to any other scheme; otherwise, the optimal moderation involves moderating the low-reputation group as little as possible (i.e., $\alpha_0 = 0$ and $\alpha_1 = M(0|w)$ if $M(0|w) < 1$; otherwise, $\alpha_0 = M^{-1}(1|w)$ and $\alpha_1 = 1$). In particular, when $\mu s \geq 1/2$, equally moderating all commentators is superior to any other schemes.

When moderators pay more attention to low-reputation commentators, high-reputation opportunistic commentators have less incentive to exert effort and might free ride. When the moderation probabilities are not high enough (to induce high-reputation opportunistic commentators to exert effort), reducing the moderation on high-reputation commentators results in higher value for the commentators because of the increased chance of free riding. Meanwhile, reducing the moderation on low-reputation commentators can increase the value of staying at low reputation (by considering the value from exerting no effort, given the commentators' indifference). Therefore, properly reducing moderation on high-reputation commentators and reducing it on low-reputation commentators can keep opportunistic commentators' incentives unchanged, and thus keep the system performance unchanged. Therefore, the most cost-effective approach is to moderate low reputation at the minimum probability for a certain level of performance. The line $\alpha_1 = H_1(\alpha_0|w)$ in Figure 1 illustrates such intuition: as the slope of the line is positive, any point with smaller α_0 and α_1 can lead to the same system performance as other points while requiring less moderation resources; thus, the most cost-effective approach is to not moderate high reputations at all. When the moderation probabilities are high enough to induce some high-reputation opportunistic commentators to exert effort, decreasing the moderation probability on low-reputation commentators decreases the value for them because of the lower chance of receiving maximum readership if $k \geq s$. To decrease the value for high-reputation commentators, we need to reduce the moderation probability for them, using the same argument for the case with $\alpha_1 > \alpha_0$. This explains why the slope of $\alpha_1 = H_2(\alpha_0|w)$ in Figure 1 is positive, and thus moderating high and low reputations equally is most cost-effective.

When moderators pay more attention to high-reputation commentators, high-reputation opportunistic commentators have greater incentive to exert effort than their low-reputation counterparts. For cases in which some opportunistic commentators exert effort, it must be that high-reputation opportunistic commentators exert effort and low-reputation opportunistic ones do not. In these cases, increasing moderation probability on low-reputation commentators (resulting in less free riding) lowers the value of staying at low reputation. In contrast, to lower the high-reputation value we need to reduce the moderation probability on high reputation, considering that high-reputation opportunistic commentators benefit from early moderation (by receiving maximum readership) if their performance with effort is better than that of dedicated commentators (i.e., $k \geq s$). Therefore, properly increasing moderation on commentators of low reputation and decreasing that on commentators of high reputation can keep opportunistic commentators' incentives unchanged, and thus keep the system performance unchanged, as the difference in reputation value influences commentators' incentives. In some sense, moderation on high reputation and moderation on low reputation are substitutes. This also explains why the slope of $\alpha_1 = M(\alpha_0|w)$ is negative in Figure 2. Proposition 4 characterizes the condition under which moderation of low-reputation commentators is more effective. In particular, when there is a significant proportion of dedicated commentators and their comments are of decent quality such that $\mu s \geq 1/2$ (i.e., when dedicated commentators play a significant role in a community), moderating low reputation is always more effective.

Under $\mu s \geq 1/2$, Proposition 4 predicts that it is optimal either to moderate high and low reputations with equal probability or to moderate low reputation only. With the condition specified in the proposition, we can further pinpoint the condition under which either scheme is optimal.

> *Corollary 3: To sustain a proportion w of opportunistic commentators in high reputation in a steady-state equilibrium under $k \geq s$ and $\mu s \geq 1/2$, (a) if $w \leq k/(1 + k)$, it is optimal to moderate low reputations only, and (b) if $w > k/(1 + k)$, it is optimal to moderate all commentators equally.*

The reason for the optimality is as follows: when $w \leq k/(1 + k)$, according to Proposition 4, for all (α_0, α_1) with $\alpha_1 \geq \alpha_0$, moderating commentators equally (i.e., $\alpha_0 = \alpha_1$) is the best choice. Meanwhile, $\alpha_0 = \alpha_1$ is dominated by the optimal scheme among all (α_0, α_1) with $\alpha_1 \leq \alpha_0$—moderating low reputation only (i.e., $\alpha_1 = 0$), which is thus the optimal among all the possible schemes. When $w > k/(1 + k)$, equally moderating commentators is unambiguously optimal for all (α_0, α_1) with $\alpha_1 \geq \alpha_0$ and for all (α_0, α_1) with $\alpha_1 \leq \alpha_0$.

Based on Corollary 3 and noting $n_1 r_1 + n_0 r_0 = \mu s + (1 - \mu)w$, the optimization problem in Equation (23) under $k \geq s$ and $\mu s \geq 1/2$ reduces to

$$\max\left\{\max_{0\leq w\leq \frac{k}{1+k}}\left[\mu s+(1-\mu)w\right]-C\left(n_0(w)\alpha_0(w)\right),\ \max_{\frac{k}{1+k}<w\leq k}\left[\mu s+(1-\mu)w\right]-C\left(\alpha_0(w)\right)\right\},$$

where the first $\alpha_0(w)$ is determined by $H_1(\alpha_0|w) = \alpha_1 = 0$ and the second $\alpha_0(w)$ is determined by $H_2(\alpha_0|w) = \alpha_1 = \alpha_0$.

The following example illustrates how moderating low-reputation commentators only might indeed appear to be the optimal solution:

> *Example: Let $\mu = 3/4$, $s = 2/3$, $c = 1/2$, $\beta = 4/5$, $k = 1$, and specify the cost function as $2(n_1\alpha_1 + n_0\alpha_0)^2$. We can verify that simply moderating low-reputation commentators only with probability 1/2 can yield a net value of 0.55, whereas the best result with an equal moderation probability is to moderate nobody, which yields a net value of 1/2.*

In fact, the relative size of the population of dedicated commentators to that of the opportunistic ones plays an important role in the choice of optimal moderation. When dedicated commentators are the majority and most of them are of high reputation, as in the example given, moderating high reputation becomes very costly without much benefit because those dedicated commentators contribute anyway. In contrast, moderating low-reputation commentators is less costly because of the relatively small population there, and properly imposing some moderation might motivate some opportunistic commentators to exert effort.

Connecting this observation to Proposition 3 and Corollary 2, we conclude that the interesting and seemingly abnormal results on reputation oscillation and reverse reputation can arise as an optimal solution.

In some special cases, the total moderation resource (say, α_T) is exogenously given, or the moderation cost is extremely high beyond a certain level in light of the cost function we have already discussed. According to the above proposition, if dedicated commentators play a significant role in a community, it is optimal either to moderate commentators with equal probability (i.e., *even moderation*) or to moderate the low-reputation group only. If the moderation probability is low (lower than α_L), even moderation is unable to induce opportunistic commentators to exert effort. In contrast, if the same moderation resource is applied to the low-reputation group, it might provide incentive to some commentators in that group to exert effort. So we have the following result:

> *Corollary 4: When the total moderation α_T is exogenously given and less than α_L, if $k \geq s$ and $\mu s \geq 1/2$, moderating the low-reputation group only is optimal.*

In general, when the moderation resource is limited, equal allocation of the moderation resource dilutes the moderation frequency and thus dilutes the opportunistic commentators' incentive to exert effort. As a result, opportunistic commentators might exert no effort in either reputation category. In contrast, by concentrating the resources on one reputation category, enough incentive for the opportunistic commentators with that reputation to exert effort might be provided because of the increased current-stage payoff.

In distributed moderation systems such as that in Slashdot, moderation is performed not by a central moderator but by distributed ones, as mentioned. In these instances, we advise system designers to provide detailed moderation guidance for potential moderators. For example, designers should tell them to focus more on high reputation, or vice versa. Furthermore, the guidance should depend on components of the commentator population and should be adjusted accordingly as the population changes.

Extensions and Discussion

So far, we have assumed that the moderation is perfect in the sense that it always correctly judges comments and that each comment is certainly moderated by the end of the period. In general, relaxing these assumptions affects opportunistic commentators' incentives. However, the intuition of our main results holds. In this section, we briefly discuss two cases by relaxing each of these two assumptions: imperfect moderation and probabilistic moderation. We also provide some discussion on the reputation measure.

Imperfect Moderation

In general, moderation cannot be perfect because of, for example, the limit of moderators' knowledge or even moderators' operational mistakes. For illustration, we

assume that moderators fairly judge a high-quality comment with probability p and always recognize low-quality comments as low quality. Then the payoff functions in Equation (3) can be reformulated as[6]

$$v_i = \max_{e\in\{0,1\}} \left[\alpha ekp + (1-\alpha) r_i\right] + \beta\left[ekpv_1 + (1-ekp)v_0\right] - ce, \text{ for } i \in \{0,1\}. \tag{26}$$

As in Equations (8) and (11), we can derive α_H and α_L under imperfect moderation as

$$\begin{aligned} \alpha_H &= \frac{\frac{c}{kp} - \beta(r_1 - r_0)}{1-\beta(r_1 - r_0)} \\ \alpha_L &= \frac{\frac{c}{kp} - \beta(s - r_0)}{1-\beta(s - r_0)}, \end{aligned} \tag{27}$$

where r_i, $i \in \{0, 1\}$, can be obtained in a similar way (see the Appendix). We assume $c < kp$.

Proposition 5 (Imperfect Moderation): Under any $\alpha \geq \alpha_H$, exerting effort can be sustained as an equilibrium. Under any $\alpha \leq \alpha_L$, exerting no effort can be sustained as an equilibrium. Both α_H and α_L decrease in p.

Moderation provides opportunistic commentators with incentive to exert effort because otherwise they could get caught and derive nothing but a low reputation for the next period. Intuitively, as the quality of moderation increases, commentators are more motivated because they are more likely to be fairly judged. Note that α_H measures the lower bound of moderation frequency to induce opportunistic commentators' effort exertion under both reputation values. So, under a higher-quality moderation, a relatively lower moderation frequency could achieve the same goal of inducing effort. Similar intuition holds for the decrease of α_L as p increases.

Proposition 5 implies that it is beneficial for communities to improve the quality of moderation. Moderators could make mistakes because of their knowledge limitations or misunderstandings. In this sense, clearly stating the community mission to commentators and especially to moderators is critical. For example, what is the purpose of the online community? What kinds of comments (e.g., informative) are encouraged and what (e.g., off topic) are discouraged? Mismoderation can also occur because moderators may have their own agenda, such as a commercial purpose. For example, on Slashdot, most moderation is performed by moderators who are randomly selected from the commentator pool. These moderators could be opportunistic or strategic when they moderate comments. To enhance the moderation system, Slashdot introduced a meta-moderation mechanism, in which moderation or moderators' judgment on the quality of comments is judged as fair or unfair by another group of users. Meta-moderation results affect moderators' reputations. Therefore, the meta-moderation not only can correct moderators' misjudgment to some extent but also can motivate opportunistic moderators to perform moderation fairly. Implementing another level

of moderation, meaning moderation on moderators, is an effective way to ensure the quality of moderation.

Probabilistic Moderation

We can also relax the assumption that each comment is certainly moderated by the end of the period by assuming instead that comments are moderated (by early moderation or late moderation) with probability θ. In addition, conditional on being moderated, a comment is moderated by early moderation with probability α as in the baseline model. If a comment does not get moderated by the end of the period, the commentator's reputation stays unchanged. Then the payoff functions in Equation (3) can be reformulated as

$$v_i = \max_{e\in\{0,1\}} \left[\alpha\theta ek + (1-\alpha\theta) r_i\right] + \beta\left[\theta\left(ekv_1 + (1-ek)v_0\right) + (1-\theta)v_i\right] - ce. \quad (28)$$

We can conduct the same analysis to obtain similar results as those in our baseline model. For example, as in Equation (8), we can derive α_H under probabilistic moderation as

$$\alpha_H = \frac{\dfrac{c}{k\theta} - \dfrac{\beta}{1-\beta(1-\theta)}(r_1 - r_0)}{1 - \dfrac{\beta\theta}{1-\beta(1-\theta)}(r_1 - r_0)},$$

where r_i, $i \in \{0,1\}$, can be obtained in a similar way (see the Appendix). We assume $c < k\theta$. Similar to the baseline model, under any $\alpha \geq \alpha_H$, exerting effort can be sustained as an equilibrium.

A More General Reputation Measure

So far, we have assumed that commentators' reputations are simply based on their performance of the latest period. Such an assumption is mainly for tractability. The reputation measure in the real world could be more complicated. For example, on Slashdot, each commentator has a reputation score in the system, which increases (decreases) by 1 if the comment of the current period is judged as high (low) quality. The reputation score is capped at some value (i.e., 50), which is to "keep people from running up insane karma scores, and then being immune from moderation" (http://slashdot.org/faq/com-mod.shtml). Similarly, there is a minimum reputation score that commentators can get (i.e., –25), and thus the score ranges at a certain interval (from –25 to 50). Commentators' karma or reputation labels (e.g., terrible, bad, positive, excellent) are based on their underlying reputation scores. Thus, a comment's default score essentially depends on the commentator's underlying reputation score.

We can consider a more general reputation measure, along the line of Slashdot's practice. For example, each commentator has a reputation score in the system. The score can range from L to N, with L being the lowest score and N the highest. If the

most recent comment is judged to be high (low) quality, the reputation score increases (decreases) by 1. The reputation displayed to users has two tiers: high and low. Without loss of generality, we assume that if a commentator's reputation score is greater than 0, he or she is in a high-reputation tier and otherwise is in low reputation. The case we discussed in the baseline model is a special case of this reputation measure with $N = 1$ and $L = 0$. We can then formulate the payoff functions for opportunistic commentators as

$$v_i = \max_{e\in\{0,1\}} \left[\alpha ek + (1-\alpha) r_j\right] + \beta\left[ekv_{i+1} + (1-ek)v_{i-1}\right] - ce, \text{ for } i \in \{L,...,N\},$$

where $r_j = r_1$ if $i > 0$ and $r_j = r_0$ otherwise, and $v_{N+1} \equiv v_N$ and $v_{L-1} \equiv v_L$ for notation simplicity.

Under this more general reputation measure, we expect that many insights derived from the baseline model continue to hold. We next use the case with $N = 1$ and $L = -1$ to illustrate the existence of α_H as in the baseline model. The balance between the marginal benefit and the marginal cost for commentators with reputation score i is again captured by the first-order derivative of the payoff function $\alpha k + \beta k(v_{i+1} - v_{i-1}) - c$, as before.

For the equilibrium with all opportunistic commentators exerting effort, their expected payoffs in equilibrium are

$$\begin{aligned} v_1 &= \alpha k + (1-\alpha) r_1 + \beta\left[kv_1 + (1-k)v_0\right] - c \\ v_0 &= \alpha k + (1-\alpha) r_0 + \beta\left[kv_1 + (1-k)v_{-1}\right] - c \\ v_{-1} &= \alpha k + (1-\alpha) r_0 + \beta\left[kv_0 + (1-k)v_{-1}\right] - c. \end{aligned}$$

Similar to Equation (8), we can derive α_H under this reputation measure:

$$\alpha_H = \frac{c/k - \dfrac{\beta^2 k}{1-\beta^2 k(1-k)}(r_1 - r_0)}{1 - \dfrac{\beta^2 k}{1-\beta^2 k(1-k)}(r_1 - r_0)},$$

where r_i, $i \in \{0, 1\}$, can be obtained in a similar way (see the Appendix). Under any $\alpha \geq \alpha_H$, exerting effort can be sustained as an equilibrium.

Conclusion

In this paper, we investigated the effect of a moderation system on the performance of online communities. We first considered equal moderation probability for different reputations and found that moderation probabilities critically affect opportunistic commentators' behavior. In particular, there is a lower bound on the moderation probability to induce effort and an upper bound to induce no effort. With a reputation system without moderation viewed as a benchmark, we showed that the reputation system with moderation always outperforms the benchmark system. We then studied a model with differentiated moderation probabilities for different reputations, where

we discovered reverse reputation and reputation oscillation. It was shown that agents in the low-reputation category might exert more effort than those in the high-reputation category and then exploit their reputation when they reach the high levels. As a result, the expected performance from the low-reputation category is even better than that from the high-reputation one. Finally, we discussed the optimal moderation resource allocation. We found that when moderation is costly, optimal moderation involves moderating commentators with equal probability or moderating low-reputation commentators under some mild conditions. We also illustrated that reputation oscillation and reverse reputation can arise in equilibrium, even under the optimal moderation allocation.

Our study provides insights for online community governance. For the purpose of inducing quality content, an online community should introduce a moderation system to monitor commentator-generated content. Promotional chats are commonly observed over the Internet [12]. Moderation not only effectively screens out this biased information but also regulates the advertisers or other commentators who otherwise would easily take advantage of the anonymity in the communities. Also, it is worth noting that the frequency of moderation is critical and should be properly chosen for optimal performance of the online community. For instance, when dedicated commentators play a significant role in an online community and opportunistic commentators are capable of generating as high-quality content as dedicated commentators, the best allocation of moderation resources is either to equally moderate both reputation groups or to moderate the low-reputation group only. If resources are adequate to induce most of the opportunistic commentators to generate high-quality content, both high- and low-reputation groups should be equally monitored. If moderation resources are limited (e.g., in terms of personnel and system capacity), resources should be directed toward the low-reputation group.

The current study can be extended in several directions. First, it is sensible to introduce an adverse selection problem with opportunistic commentators. In general, opportunistic commentators can differ in their hidden abilities to generate high-quality information. These hidden abilities can result from various factors, such as their knowledge level and the opportunity cost of their effort/time. As a result, a simple reputation measure that considers only the recent moderation outcome is insufficient because the reputation not only is about the threat of future punishment but also involves learning about agents' types. A model with an adverse selection problem can be expected to offer more significant results. In addition, once a richer reputation measure is introduced, the moderation scheme can then be further refined based on agents' reputation/history. How to tailor moderation for commentators with different reputations is another question for future research.

Second, it is interesting and important to consider competition among online communities for the next step. Competition is ubiquitous on the Internet. While our paper derives insight about the effect and design of moderation policy isolated from the competition, studying the moderation mechanism in a competitive environment can complement and enrich our results.

Finally, conducting empirical or experimental tests on our results would be another interesting research direction. The results derived from our analytical model provide

many testable predictions about the effect of moderation policy. For example, high moderation probability generally leads to a higher quality of comments, and reverse reputation may occur when the moderation probability on low-reputation commentators is higher than on high-reputation ones. To test such predictions, one might use the quality of comments on Slashdot, which can be readily measured because the score that a comment receives reflects its quality; the moderation probability can be approximated in some way as well (e.g., by using the percentage of comments being moderated to the total comments in each category).

Acknowledgments: The authors thank the three anonymous reviewers as well as Zoltn Hidvgi and Barrie R. Nault for their generous inputs to and valuable comments on this research. They thank Rob "Commander Taco" Malda (the founder of Slashdot) for his insights on moderating online communities and moderation systems. They also thank participants at the Twenty-Seventh International Conference on Information Systems, INFORMS International 2009, the Third China Workshop on Information Management, and INFORMS 2009 for their helpful feedback. The authors gratefully acknowledge financial support from the NET Institute (www.netinst.org). They thank the Social Sciences and Humanities Research Council of Canada for support.

Notes

1. See "Why Wikipedia Must Jettison Its Anti-Elitism," at www.kuro5hin.org/story/2004/12/30/142458/25/.

2. Provided a linear cost structure, the analysis of the case with continuous effort level e, $e \in [0, 1]$, is equivalent to the analysis we conduct under the current setting.

3. Technically speaking, $\alpha_L > \alpha_H$ may occur. In that case, multiple equilibria exist for a certain range of moderation probabilities.

4. In fact, under moderation probability in this range, asymmetric equilibria may exist. For example, it could be that high-reputation opportunistic commentators have a higher probability to exert effort in some equilibria.

5. Note that the following equilibrium in which low-reputation opportunistic commentators exert effort with a higher probability than their high-reputation counterparts can also be sustained under $\alpha_0 = \alpha_1$. In other words, the reputation oscillation equilibrium also exists under the same moderation probability as under the differentiated moderation probabilities, although the reasons differ. The former is because opportunistic commentators are indifferent in exerting effort or not (and the described equilibrium is one *asymmetric* equilibrium mentioned in note 4), whereas the latter is because opportunistic commentators have greater incentive to exert effort under high reputation.

6. For simplicity, we continue to assume that the moderated low-quality comments from early moderation get 0 readership, although some of those comments may be of high quality.

References

1. Ba, S., and Pavlou, P.A. Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quarterly, 26,* 3 (2002), 243–268.

2. Benabou, R., and Tirole, J. Incentives and prosocial behavior. *American Economic Review, 96,* 5 (2006), 1652–1678.

3. Chromatic, A.B., and Krieger, D. *Running Weblogs with Slash.* Sebastopol, CA: O'Reilly Media, 2002.

4. Dellarocas, C. Reputation mechanism design in online trading environments with pure moral hazard. *Information Systems Research, 16,* 2 (2005), 209–230.

5. DeLone, W.H., and McLean, E.R. The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems, 19,* 4 (Spring 2003), 9–30.

6. Fang, Y., and Neufeld, D. Understanding sustained participation in open source software projects. *Journal of Management Information Systems, 25,* 4 (Spring 2009), 9–50.

7. Forman, C.; Ghose, A.; and Wiesenfeld, B. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research, 19,* 3 (2008), 291–313.

8. Forte, A.; Larco, V.; and Bruckman, A. Decentralization in wikipedia governance. *Journal of Management Information Systems, 26,* 1 (Summer 2009), 49–72.

9. Internet grows up. *Financial Times* (August 28, 2009) (available at www.ft.com/cms/s/0/6928e7fc-9406-11de-9c57-00144feabdc0.html#axzz1ZpicC0E8/).

10. Lampe, C., and Resnick, P. Slash(dot) and burn: Distributed moderation in a large online conversation space. Paper presented at the ACM CHI 2004 Conference on Human Factors in Computing Systems, Vienna, Austria, April 24–29, 2004.

11. Lopes, A.B., and Galletta, D.F. Consumer perceptions and willingness to pay for intrinsically motivated online content. *Journal of Management Information Systems, 23,* 2 (Fall 2006), 203–231.

12. Mayzlin, D. Promotional chat on the Internet. *Marketing Science, 25,* 2 (March–April 2006), 155–163.

13. Mookherjee, D., and Png, I. Optimal auditing, insurance, and redistribution. *Quarterly Journal of Economics, 104,* 2 (1989), 399–415.

14. Susarla, A.; Tan, Y.; and Oh, J. Social networks and the diffusion of user-generated content: Evidence from YouTube. *Information Systems Research,* forthcoming.

15. Wasko, M.M., and Faraj, S. Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly, 29,* 1 (2005), 35–57.

16. Wattal, S.; Racherla, P.; and Mandviwalla, M. Network externalities and technology use: A quantitative analysis of intraorganizational blogs. *Journal of Management Information Systems, 27,* 1 (Summer 2010), 145–174.

17. Zhang, X., and Zhu, F. Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review, 101,* 4 (2011), 1601–1615.

18. Zhu, F., and Zhang, X. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74,* 2 (2010), 133–148.

Appendix

Proof of Proposition 2

FOR THE CASE WHERE OPPORTUNISTIC COMMENTATORS ADOPT A MIXED STRATEGY, note that the expected quality is the weighted average of the success rate s and mk. The ratio of weight for s to mk determines the value. If $s > mk$, the expected quality is increasing in the ratio of weight for s to that for mk. So, to conclude that $r_1 > r_0$ is equivalent to showing

$$\frac{\mu s}{(1-\mu)mk} \geq \frac{\mu(1-s)}{(1-\mu)(1-mk)}, \tag{A1}$$

which can be verified as true. If $s < mk$, similarly, we can derive $r_1 > r_0$. $s = mk$ is a special case in which $r_1 = r_0$.

In the equilibrium with effort, it can be similarly shown that $r_1 > r_0$. In the equilibrium with no effort, it is easy to see $r_1 = s > r_0$.

Proof of Proposition 3

(a) For $w \in [0, k/(1+k))$, incentive conditions require that $\alpha_1 k + \beta k(v_1 - v_0) - c < 0$ and $\alpha_0 k + \beta k(v_1 - v_0) - c = 0$. Because $\alpha_1 < \alpha_0$, we need to check only the second condition. Notice that we have $v_1 = (1-\alpha_1)r_1(w) + \beta v_0$ and $v_0 = (1-\alpha_0)r_0(w) + \beta v_0$. So incentive conditions lead to

$$(\alpha_0 k - c) + \beta k\left[(1-\alpha_1)r_1(w) - (1-\alpha_0)r_0(w)\right] = 0. \tag{A2}$$

Therefore,

$$\begin{aligned}\alpha_1 &= \frac{\left[1+\beta r_0(w)\right]\alpha_0 + \beta\left[r_1(w) - r_0(w)\right] - c/k}{\beta r_1(w)} = H_1(\alpha_0 | w) \\ &= 1 - \frac{\beta(1-\alpha_0)r_0(w) - \alpha_0 + c/k}{\beta r_1(w)}.\end{aligned} \tag{A3}$$

Because $r_1(w)$ decreases in w and $r_0(w)$ increases in w, the right-hand side of Equation (A3) is decreasing in w. So, if

$$H_1\left(\alpha_0 \left| \frac{k}{1+k}\right.\right) < \alpha_1 \le H_1(\alpha_0 | 0),$$

we can get a unique solution to $\alpha_1 = H_1(\alpha_0 | w)$.

(b) For $w \in (k/(1+k), k]$, incentive conditions require that $\alpha_1 k + \beta k(v_1 - v_0) - c = 0$. Noticing that $v_1 - v_0 = (\alpha_1 - \alpha_0)k + (1-\alpha_1)r_1 - (1-\alpha_0)r_0$, we can derive

$$\alpha_1 = H_2(\alpha_0 | w) \equiv \frac{\beta\left[k - r_0(w)\right]\alpha_0 + \beta\left[r_0(w) - r_1(w)\right] + c/k}{1 + \beta k - \beta r_1(w)}.$$

Similarly, if

$$H_2\left(\alpha_0 \left| \frac{k}{1+k}\right.\right) < \alpha_1 \le H_2(\alpha_0 | k),$$

there is α_1 such that $\alpha_1 = H_2(\alpha_0 | w)$.

(c) In the case with $w = k/(1+k)$, incentive conditions require that $\alpha_1 k + \beta k(v_1 - v_0) - c \le 0$ and $\alpha_0 k + \beta k(v_1 - v_0) - c \ge 0$. From the proofs in (a) and (b), we can verify that the incentive conditions hold when $\alpha_1 \le H_1(\alpha_0 | k/(1+k))$ and $\alpha_1 \le H_2(\alpha_0 | k/(1+k))$.

Proof of Corollary 2

For $w \in [0, k/(1+k)]$, based on Equations (15) and (16), the condition for $r_1(w) < r_0(w)$ is

$$\mu s s(1-\mu)(1-w) < \mu s(1-\mu)w + (1-\mu)w\left[\mu(1-s)s + (1-\mu)w\right],$$

which can be reduced to $(1-\mu)w^2 + 2\mu s w - \mu s^2 > 0$ by simple algebra, or, equivalently,

$$w > \frac{\sqrt{\mu} - \mu}{1 - \mu} s = \frac{\sqrt{\mu}}{1 + \sqrt{\mu}} s.$$

Proof of Proposition 4

(a) Under (α_0, α_1) with $\alpha_1 \le \alpha_0$, for $w = 0$, the minimum resource required is trivially 0. For any $w \in (0, k/(1+k))$, by Proposition 3, $\alpha_1 = H_1(\alpha_0 | w)$ defines a line on which any pair (α_0, α_1) yields the same proportion w in equilibrium. Given the positive slope of $\alpha_1 = H_1(\alpha_0 | w)$, the coefficient of α_0 in Equation (24) is positive when α_1 is substituted in. Therefore, the optimal solution is the minimum α_0 possible on $\alpha_1 = H_1(\alpha_0 | w)$, which is $\alpha_1 = 0$ and α_0 determined by $0 = H_1(\alpha_0 | w)$. For $w \in (k/(1+k)/k)$, under $k \ge s$, $k \ge r_0(w)$ and thus $\alpha_1 = H_2(\alpha_0 | w)$ has a positive slope. Substituting $\alpha_1 = H(\alpha_0 | w)$ into Equation (24), the coefficient of α_0 is positive, and therefore the optimal solution is the minimum α_0 possible on $\alpha_1 = H_2(\alpha_0 | w)$, which is $\alpha_1 = \alpha_0$. By Figure 1, for $w = k/(1+k)$, the optimal solution is clearly $\alpha_1 = 0$, and α_0 is determined by $0 = H_1(\alpha_0 | k/(1 + \mathrm{k}))$. For $w = k$, the optimal solution is $\alpha_1 = \alpha_0 = H_2(\alpha_0 | k)$.

(b) Under (α_0, α_1) with $\alpha_1 \ge \alpha_0$, for $w = 0$, the minimum resource required is trivially 0. For $w \in (0, k)$, under $k \ge s$, we have $k > r_1(w)$, and $\alpha_1 = M(\alpha_0 | w)$ in Equation (22) has a negative slope. Substituting $\alpha_1 = M(\alpha_0 | w)$ into Equation (24), the coefficient of α_0 is

$$\frac{-\left(1 - \beta k + \beta r_0(w)\right)}{\beta\left(k - r_1(w)\right)} n_1 + n_0.$$

The sign of the coefficient is the same as

$$\begin{aligned} -\frac{1}{\beta}\left(1 - \beta k + \beta r_0\right) n_1 + n_0\left(k - r_1\right) &= k - \frac{1}{\beta} n_1 - \left(r_0 n_1 + r_1 n_0\right) \\ &= k - \left(\frac{1}{\beta} - 1\right) n_1 - \left(r_0 + r_1\right), \end{aligned} \tag{A4}$$

where the second equality is because $r_0 n_0 + r_1 n_1 = n_1$ (i.e., the proportion of commentators with high reputation is constant), and thus $r_0 + r_1 = (r_0 + r_1)(n_0 + n_1) = (r_0 n_1 + r_1 n_0) + n_1$. Notice that $n_1 = \mu s + (1 - \mu) w$ is increasing in w, and r_1 and r_0 are increasing in w according to Equations (19) and (20). Therefore, the term in the right-hand side of Equation (A4) is decreasing in w. When $w > w^*$, the term in Equation (A4) and also the coefficient of α_0 are negative, and thus the optimal solution is the maximum α_0 possible on $\alpha_1 = M(\alpha_0 | w)$, which is $\alpha_0 = \alpha_1$. When $w < w^*$, the coefficient of α_0 is positive and thus the optimal solution is the minimum α_0 possible on $\alpha_1 = M(\alpha_0 | w)$, which leads to either $\alpha_0 = 0$ and $\alpha_1 = M(0 | w)$ (if $M(0 | w) \le 1$) or $\alpha_1 = 1$ and $\alpha_0 = M^{-1}(1 | w)$.

If $\mu s \ge 1/2$, then

$$r_0 n_1 + r_1 n_0 = \frac{n_1}{n_0} r_0 n_0 + \frac{n_0}{n_1}\left(r_1 n_1 - n_1 + n_1\right) = n_0 + \left(\frac{n_1}{n_0} - \frac{n_0}{n_1}\right) r_0 n_0 > n_0,$$

where the second equality is because $r_0 n_0 + r_1 n_1 = n_1$ and the inequality is because $n_1 \geq n_0$. Also,

$$k - \frac{1}{\beta} n_1 < 1 - n_1 = n_0.$$

Therefore, the term in Equation (A4), and thus the coefficient of α_0, are negative for all w. Therefore, the optimal solution is the maximum α_0 possible on $\alpha_1 = M(\alpha_0 | w)$, which is $\alpha_0 = \alpha_1$.

Proof of Proposition 5

We can reorganize α_H as follows:

$$\alpha_H = \frac{\frac{c}{kp} - \beta(r_1 - r_0)}{1 - \beta(r_1 - r_0)} = 1 - \frac{1 - \frac{c}{kp}}{1 - \beta(r_1 - r_0)}, \tag{A5}$$

where

$$\begin{aligned} r_1 &= \frac{\mu sps + (1-\mu)kpk}{\mu sp + (1-\mu)kp} \\ r_0 &= \frac{\mu(1-sp)s + (1-\mu)(1-kp)k}{\mu(1-sp) + (1-\mu)(1-kp)}. \end{aligned} \tag{A6}$$

Notice that r_0 is the weighted average of the expected success rates s (from dedicated commentators) and k (from opportunistic commentators). So r_0 increases (decreases) in $(1 - \mu)(1 - kp)/(\mu(1 - sp))$, the weight ratio of k and s, if $s < k$ (if $s > k$). It is easy to show that $(1 - \mu)(1 - kp)/(\mu(1 - sp))$ decreases (increases) in p if $s < k$ (if $s > k$), so r_0 decreases in p. Also, notice that r_1 is independent of p. Therefore the second term on the right-hand side of (A5) increases in p because the numerator increases in p and the denominator decreases in p. Thus, α_H decreases in p.

Similarly, we can show that α_L is decreasing in p.

Derivation of α_H Under Probabilistic Moderation

The first-order derivative of the payoff is $\alpha\theta k + \beta\theta k(v_1 - v_0) - c$. The opportunistic commentators' expected payoff in an equilibrium with effort is

$$v_i = \left[\alpha\theta k + (1 - \alpha\theta) r_i\right] + \beta\left[\theta\left(kv_1 + (1-k)v_0\right) + (1-\theta)v_i\right] - c.$$

Therefore, we have $v_1 - v_0 = (1 - \alpha\theta)(r_1 - r_0) + \beta(1 - \theta)(v_1 - v_0)$ or

$$v_1 - v_0 = \frac{1 - \alpha\theta}{1 - \beta(1-\theta)}(r_1 - r_0).$$

Suppose we have y dedicated commentators in high reputation. According to the steady-state condition, we have

$$y\left[\theta s+(1-\theta)\right]+(\mu-y)\theta s=y.$$

Thus, $y=\mu s$. Similarly, we have $(1-\mu)k$ opportunistic commentators in high reputation. Therefore,

$$r_1=\frac{\mu ss+(1-\mu)kk}{\mu s+(1-\mu)k}$$

$$r_0=\frac{\mu(1-s)s+(1-\mu)(1-k)k}{\mu(1-s)+(1-\mu)(1-k)}.$$

Substituting v_1-v_0 into the above first-order derivative, and making it zero, we can derive

$$\alpha_H=\frac{\frac{c}{k\theta}-\frac{\beta}{1-\beta(1-\theta)}(r_1-r_0)}{1-\frac{\beta\theta}{1-\beta(1-\theta)}(r_1-r_0)}.$$

Derivation of α_H Under the More General Reputation Measure

For the equilibrium with effort, we have $v_1-v_0=(1-\alpha)(r_1-r_0)+\beta(1-k)(v_0-v_{-1})$ and $v_0-v_{-1}=\beta k(v_1-v_0)$. Therefore,

$$v_1-v_0=\frac{(1-\alpha)(r_1-r_0)}{1-\beta^2k(1-k)}.$$

For dedicated commentators, suppose x of them are at score 1, y of them are at score 0, and $\mu-x-y$ of them are at score –1. According to the steady-state condition, we have $(x+y)s=x$ and $x(1-s)+(\mu-x-y)s=y$. We can then derive

$$x=\frac{\mu s}{s^2+2(1-s)}.$$

Similarly, we define x' as the proportion of opportunistic commentators at score 1, and

$$x'=\frac{(1-\mu)k}{k^2+2(1-k)}.$$

Then we can formulate r_1 and r_0 as

$$r_1=\frac{\mu xs+(1-\mu)x'k}{\mu x+(1-\mu)x'}$$

$$r_0 = \frac{\mu(1-x)s + (1-\mu)(1-x')k}{\mu(1-x) + (1-\mu)(1-x')}.$$

Also notice that $v_0 - v_{-1} < v_1 - v_0 < v_1 - v_{-1}$. We can rearrange the first-order incentive condition as

$$\alpha k + \beta k(v_0 - v_{-1}) - c \geq 0.$$

We can derive α_H by binding the above inequality and substituting in $(v_0 - v_{-1})$.