Working Paper #09-11

September 2009

Moderated Online Communities and User-Generated Content

Jianqing Chen
University of Calgary

Hong Xu
University of Texas at Austin

Andrew B. Whinston
University of Texas at Austin

Moderated Online Communities and User-Generated Content*

Jianqing Chen
University of Calgary
jiachen@ucalgary.ca

Hong Xu
University of Texas at Austin
hong.xu@phd.mccombs.utexas.edu

Andrew B. Whinston
University of Texas at Austin
abw@uts.cc.utexas.edu

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, however, remains a challenge due to the anonymity of commentators. This paper introduces moderation into reputation systems. We show that moderation directly impacts strategic commentators incentive to generate useful information, and moderation is generally desirable to improve information quality. Interestingly, we find that when being moderated with different probabilities based on their reputations, commentators may display a pattern of reputation oscillation, in which they generate useful content to build up high reputation and then exploit their reputation. As a result, the expected performance from high-reputation commentators can be inferior to that from low-reputation ones (reversed reputation). We then investigate the optimal moderation resource allocation, and conclude that the seemingly abnormal reversed reputation could arise as an optimal result. The paper concludes with a discussion of the development of a scientific moderation system with application to academic publishing.

Keywords: moderation, reputation, online community, knowledge management

*We thank Barrie R. Nault for his valuable comments on this research. We also thank participants at 27th International Conference on Information Systems for their helpful feedback. This research is partially funded by the NET Institute whose financial support is gratefully acknowledged.

1 Introduction

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 (Parameswaran and Whinston, 2007). A large number of applications have emerged that facilitate collective actions for content generation and knowledge sharing. Examples include blogs, online product reviews, wiki applications such as Wikipedia.com, and online forums such as slashdot.org. Due to the anonymity of Internet users, however, how to ensure information quality or induce quality content remains a challenge. To deal with that, 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 "allow[ing] visitors to add, remove, edit and change content."[1] In online product reviews, like the ones on Amazon.com, any user can post reviews on any item, even if he or she has not bought it on Amazon. Online forums such as Slashdot are another example. Slashdot, a website 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 the comment under a selected topic.

As these applications have gained popularity and importance, the quality of content has become a concern. In Wikipedia, readers may be provided with content that is misleading or even incorrect. Product reviews on Amazon can be manipulated by sellers or book publishers to boost their products. On Slashdot, commentators may post some biased or useless comments; e.g., advertisers from hardware companies may post biased comments to promote their products. Wikipedia is still experimenting with different approaches to ensure the quality of content. As one of its co-founders pointed out, Wikipedia, as an encyclopedia, lacks both the usual review process and the respect for expertise of most encyclopedias.[2] Amazon

[1] http://en.wikipedia.org/wiki/Wiki

[2] http://www.kuro5hin.org/story/2004/12/30/142458/25

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, since the voting process itself can be manipulated.

Slashdot has constructed a reputation system and has been recognized for its quality of content, unlike many other social networks. On Slashdot, commentators develop reputations based on the quality of their past comments.[3] Each comment posted by a commentator receives a score ranging from -1 to 5, indicating the quality of the comment. Comments' default scores differ from each other according to the commentators' reputations. 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." This feedback affects the comment's score and the commentator's reputation. Slashdot selects moderators randomly from among eligible users, and limits the moderator status both in number of posts to be moderated (five) and in time (three days). This restriction ensures no moderator can have an undue effect on the system.[4]

Similar moderation processes have been adopted by other online communities, such as www.kuro5hin.org and www.plastic.com. In fact, the moderation process was introduced mainly to screen information. As stated by one of Slashdot's founders, "The purpose of moderation is to help people organize information;" it can help users "pick up hidden gems on the sandy beach of comments" (Chromatic et al., 2002). However, it seems the actual impact 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

[3] Slashdot uses "karma" points to measure commentators' reputations. Karma points or reputations are clustered into a small set of labels (e.g., terrible, bad, positive, excellent). Commentators can change their karma points by posting comments of quality or performing other community activity.

[4] To enhance the moderation system, Slashdot introduces a meta-moderation mechanism, where readers can volunteer to review the fairness of moderators' decisions. These reviews may affect a moderator's Karma, which in turn influences the chance that the moderator will be chosen as a moderator in the future.

moderation or the design of a moderation system. As a starting point, this paper examines the impact of moderation on the performance of an online community. We consider a community consisting of both dedicated and opportunistic commentators. The former behave altruistically, while the latter behave strategically. Commentators possess their reputations in the community and post their comments of different quality. A moderation system moderates each comment. The moderation result impacts 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 impact on the opportunistic types' 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 may exert no effort at all. When the level of moderation is in the middle, commentators may adopt a strategy mixing exertion and no exertion. It is also demonstrated 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, interestingly, 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 issues. We find that when the moderation resource is costly, optimal moderation involves either equally moderating all commentators or moderating low-reputation commentators only. In other words, it is never optimal to monitor high-reputation commentators more closely.

As information quality has been identified as an important factor in the success of information systems (DeLone and McLean, 1992), the quality of content is a natural concern for online communities. A large volume of the existing literature focuses on reputation systems. For example, Dellarocas (2005) has studied the reputation mechanism in eBay-like trading environments, with a focus on how mechanism parameters (e.g., a user's feedback profile) impact sellers' effort levels and market efficiency. The study on reputation in economics can be traced back several decades. In their seminal work, Kreps and Wilson (1982) and Milgrom and Roberts (1982) concluded that reputation effects arise with even a small amount of incomplete information on agents' types. Later, Cripps et al. (2004) showed that with imperfect monitoring, reputation cannot be sustained infinitely – if a long-run player stays in the game long enough, short-run players will eventually learn the long-run's true type and the game will inevitably revert to one of the static Nash equilibrium. In contrast to well-understood reputation systems, the moderation system has attracted little notice. Lampe and Resnick (2004) document some observations of the moderation practice and point out that "important challenges remain for designers of such systems." Our paper tries to build up a game-theoretic model to analyze commentators' incentives and study the impact of moderation, and aims to address why moderation works and how it can be improved.

Study of online communities also has been concerned with users' motivation for their voluntarily participation in and contribution to communities. Based on their data, Wasko and Faraj (2005) find several factors related to users' motivation to contribute, such as the perception of enhancing their professional reputations. Bateman et al. (2006) study this issue, drawing on organizational commitment theory. Benabou and Tirole (2006) 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 may be driven by intrinsic value, whereas the latter pursues monetary benefit.

The rest of the paper is organized as follows. In section 2 we lay out our model. We analyze the equilibrium effort choice under the same moderation probabilities in section 3 and under differentiated moderation probabilities in section 4. In section 5, we investigate the optimal moderation resource allocation. Some extensions and discussion are offered in section 6. Section 7 concludes the paper.

2 Model

We consider an online community in an infinite-period horizon, in which a large number of commentators post comments and develop reputations in doing so. 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 perceived 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. Dedicated types always post their true opinions and behave like altruists. This would be because they derive a great deal of intrinsic value from the community and are thus dedicated to the community posting. Due to the heterogeneity in the commentators' knowledge, some of their comments are useful, whereas others may be worthless. We assume that the proportion of useful comments is s. In contrast, opportunistic types behave strategically. They can exert effort ($e = 1$) to generate a useful comment, or exert no effort ($e = 0$) to post a worthless comment. Exerting effort incurs cost c. Cost c can be interpreted as the time commentators spend in properly organizing their opinions or investigating the topic under discussion. We normalize the population size to 1 and assume the proportion of dedicated commentators is μ, and hence the proportion of the opportunistic type is $(1 - \mu)$.

A moderation system moderates the quality of comments and labels a verified comment as useful or worthless. With probability α, a comment is moderated immediately after it is posted; otherwise, the comment is moderated at the end of the period. We term the former as early moderation or moderation, and the latter as 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, since the results are revealed at the end of the period.

A commentator may have a high reputation or a low reputation. We consider the commentator's reputation is high if the last comment is judged to be useful and low if it is deemed worthless. Such an assumption imposes little restriction, since the primary purpose of our reputation system is to deal with opportunistic types' incentive and the above simple reputation measure plays an effective sanctioning role (the threat of future punishment). In fact, as shown by Dellarocas (2005), such a reputation measure "is capable of inducing the same average levels of cooperation and total surplus as more sophisticated mechanisms."

We are interested in the impact of the moderation system on opportunistic types' behavior. We assume opportunistic types derive utility from others reading their comments. In particular, we assume the utility is linear in the size of readers. The verified useful comments from early moderation get the maximum readership, normalized to 1, and the verified worthless comments from early moderation get 0 readership. For comments with late moderation, their readership level is equal to the expected likelihood of usefulness or the expected success rate, which is also termed as the expected performance associated with those comments.

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

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

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

where β is a discount factor and r_i^t is the expected success rate of a comment from commentators with reputation i.

We will be concerned with *steady states* in which r_i^t and v_i^t are independent of time (they, of course, will 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

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

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 the above 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, since e can also be interpreted as the probability of exerting effort in our game-theoretic framework.

3 Equilibrium Performance

Moderation probabilities have a critical impact on opportunistic types' optimal choice. In this section, we investigate three cases where, in equilibrium, opportunistic types exert effort definitely, exert no effort definitely, and exert effort with some probability, respectively.

Clearly, the marginal benefit from exerting effort is the probabilistic increase in the current period payoff (α) and the increase in discounted future payoff ($\beta(v_1 - v_0)$). 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 functions (3),

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

which determines their equilibrium choice. If the above is positive, which means the marginal benefit outweighs the marginal cost, the commentator will exert effort. Otherwise, he or she 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 they possess high reputations, they also find it optimal when they possess low reputations.

Notice that dedicated types do not behave strategically, and with probability s their comments are useful regardless of their current reputations. Therefore, a proportion s of dedicated types possess high reputations.

3.1 The Equilibrium with Effort

When the probability of early moderation (*moderation probability* afterwards) is high, opportunistic types have high motivation to exert effort; otherwise, their comments would fail the early moderation and thus receive no readership. More precisely, the equilibrium with opportunistic types exerting effort requires high moderation probabilities, such that the marginal benefit outweighs the marginal cost; i.e., $\alpha + \beta(v_1 - v_0) - c \geq 0$. In such equilibria, opportunistic types exert effort and maintain their high reputations.

According to (3), the opportunistic types' expected payoffs in equilibrium are[5]

$$v_1 = (1 - \alpha)\, r_1 + \alpha - c + \beta v_1, \text{ and } v_0 = (1 - \alpha)\, r_0 + \alpha - c + \beta v_1. \tag{5}$$

The difference between the above expected payoffs, $v_1 - v_0 = (1 - \alpha)\,(r_1 - r_0)$, plays a role in determining opportunistic types' 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. We next consider nontrivial cases with $\alpha < 1$.

Recall the proportion of dedicated types with high reputations is s. Opportunistic types all have high reputations when they exert effort. So the size of the population in high reputations will be $\mu s + (1 - \mu)$, consisting of dedicated types (the first term) and opportunistic types (the second term). Since the success rates of comments are s (dedicated types) and 1 (opportunistic types), respectively, we can formulate the expected success rate or the expected performance associated with a high reputation as follows:

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

[5] We can call v_0 the equilibrium payoff, imagining that with an arbitrarily small probability ϵ the moderation makes misjudgement such that opportunistic types are still possible in low reputations even though they exert effort. For more detailed discussion on imperfect moderation, refer to Section 6.1.

For low reputations, because they are composed solely of dedicated types, the expected success rate is simply their expected performance s; i.e., $r_0 = s$. It is easy to see $r_1 > r_0$, implying that high reputations indicate higher expected performance.

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

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

Clearly, the left hand side is increasing in α. In other words, the higher the moderation probability, the more likely the opportunistic types are to exert effort. Intuitively, increasing moderation probability means increasing the chance of getting early moderation. This enhances the current period benefit from exerting effort because their comments are very likely to be revealed as valuable immediately and thus to receive the highest readership. Therefore, a higher moderation probability is more likely than a lower one to induce opportunistic types to exert effort.

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

$$\alpha_H = \frac{c - \beta \left(r_1 - s\right)}{1 - \beta \left(r_1 - s\right)}. \tag{8}$$

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, such that $c \geq 1$ (then $\alpha_H \geq 1$), no moderation scheme can induce opportunistic types to exert effort. Recall that the maximum readership/benefit that commentators can achieve is 1 at each period. Therefore, when the cost is beyond that, no opportunistic types will exert effort in any cases. For this reason, we assume that $c < 1$.

From the definition of α_H with the assumption $c < 1$, $\alpha > c$ is a sufficient condition to induce opportunistic types to exert high effort. Intuitively, $\alpha > c$ means that the expected increase in the current period payoff (α) outweighs the marginal cost c, which provides commentators with adequate incentive to exert effort.

3.2 The Equilibrium with No Effort

Because dedicated types can have high reputations and low reputations, readers have a certain expectation about their performance even for the low-reputation commentators. As a result, opportunistic types may catch a "free ride" on those dedicated types by receiving some readership while exerting no 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 opportunistic types' best choice. More precisely, when the marginal benefit from exerting effort is not enough to compensate for the marginal cost, i.e., $\alpha - c + \beta(v_1 - v_0) < 0$, opportunistic types exert no effort in equilibrium. The equilibrium expected payoff 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 types maintain low reputations because they exert no effort. As a result, the high-reputation commentators are composed purely of dedicated types, and therefore the expected performance is s (as that of dedicated types), i.e., $r_1 = s$. Low-reputation commentators consist of both dedicated types and opportunistic types. Recall that a proportion $1 - s$ of dedicated types is in the low-reputation category with those opportunistic types. We formulate the expected performance of 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

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

Clearly, the left hand side is increasing in α. In other words, the lower the moderation probability, the less likely opportunistic types 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

$$\alpha_L = \frac{c - \beta\left(s - r_0\right)}{1 - \beta\left(s - r_0\right)}. \tag{11}$$

Thus, we can derive the following lemma.

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

The intuition is as we mentioned at the beginning of this subsection. Opportunistic types 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 certain level of expectation in the performance is attributed to the dedicated types, since they always contribute, which provides opportunistic types a chance to free-ride. When the moderation probability is low and hence there is only a low chance of getting caught and ending up with nothing, opportunistic types have a strong incentive to free-ride the dedicated types. So, low moderation, no effort.

However, when the cost of effort is low enough (such that $c \leq \beta (s - r_0)$ and then $\alpha_L \leq 0$), exerting no effort cannot be sustained as an equilibrium, no matter how low the moderation probability is. This is because when free-riding is expected, the expected readership is also adjusted to a lower level in equilibrium. Meanwhile, opportunistic types always have the option to exert effort, join the high-reputation group, and obtain high expected readership. When the effort cost is very low, the benefit from free-riding will be overcome by the net benefit from exerting effort. As a result, regardless of how low the moderation probability is, opportunistic types choose to exert effort.

3.3 Mixed Strategy Equilibrium

The above analysis characterizes the opportunistic types' equilibrium effort choice when the moderation probability is very high or very low. What will be their equilibrium choice if the moderation probability is between the two, say $\alpha_L < \alpha < \alpha_H$?[6] In such cases, we can speculate that in equilibrium some opportunistic types may exert effort, whereas others do not, or they sometimes exert effort but other times do not. This involves mixed strategy equilibria.

For a mixed strategy (between exerting effort and not exerting effort) to arise in equilibrium, opportunistic types must be indifferent about exerting effort or not – otherwise

[6] Technically speaking, $\alpha_L > \alpha_H$ may occur. In that case, multi equilibria exist for a certain range of moderation probabilities.

they could always go with the more profitable option. So the marginal benefit balances the marginal cost in equilibrium, i.e., $\alpha+\beta(v_1-v_0)-c=0$. We consider a symmetric case where opportunistic types exert effort with probability m in each reputation.[7] 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; i.e., $v_1 - v_0 = (1-\alpha)(r_1 - r_0)$ (refer to (3)). The proportion of opportunistic types with high reputations will be m. Then, we can characterize the expected performance under high reputation and low reputation, respectively, as

$$\begin{aligned} r_1 &= \frac{\mu ss + (1-\mu)\, mm}{\mu s + (1-\mu)\, m}, & (12) \\ r_0 &= \frac{\mu (1-s)\, s + (1-\mu)(1-m)\, m}{\mu (1-s) + (1-\mu)(1-m)}. & (13) \end{aligned}$$

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

$$\alpha(m) = \frac{c - \beta(r_1 - r_0)}{1 - \beta(r_1 - r_0)}. \tag{14}$$

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 (14).*

A mixed strategy may arise as an equilibrium because of the externality of the benefit from free-riding. Opportunistic types benefit from pooling with or free-riding dedicated types 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 types 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 the equilibrium, the benefit from free-riding balances the net

[7] In fact, under moderation probability in this range, multiple equilibria exist. For example, it could be that proportion w of opportunistic types stay with high reputation and exert effort and the rest of opportunistic types stay with low reputation and exert no effort (notice all of them are indifferent in exerting effort or not in equilibrium).

benefit from exerting effort, which also determines the number of free-riders (the probability that opportunistic types 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 types exerting effort; For $\alpha < \alpha_L$, opportunistic types exerting no effort; For $\alpha \in [\alpha_L, \alpha_H]$, opportunistic types exerting effort with probability $m(\alpha)$ (determined by (14)).*

Since different moderation arrangements provide different incentives for opportunistic types 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 is in contrast to the case with differentiated moderation probabilities, shown in the next section.

Proposition 2 *In equilibria described above, the expected performance of high-reputation commentators is higher than that of low-reputation commentators. Formally, $r_1 > r_0$.*

This result looks very natural, since 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 types' performance. (recall that dedicated types perform at the same level under each reputation.) Notice that in each equilibrium described above, opportunistic types exert the same level of effort under each reputation because of the symmetric incentive (which is due to the same moderation probability). Therefore, the above proposition, in fact, says that the higher performance commentators dominate in the high-reputation group more than in the low-reputation group.

3.4 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 above moderation system into a pure reputation system.

Without moderation, the marginal benefit of exerting effort is from the increase in discounted future payoff ($\beta(v_1 - v_0)$) 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 proposition 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 a moderation arrangement provides a cost-effective way to implement the moderation system in online communities.

4 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 may watch low-reputation commentators more carefully, considering that they perform poorly.

In this section, we study a more general case in which the moderation system moderates comments from 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 ones α_i in the payoff function (3), we can get a similar payoff function.

The basic tradeoff in commentators' decisions remains the same, except that now we have differentiated moderation probabilities. Similar to (4), the incentive to exert effort is determined by $\alpha_i + \beta(v_1 - v_0) - c$, $i \in \{0, 1\}$. Because of the differentiated moderation probabilities, unlike the previous case, opportunistic types may choose asymmetric effort in equilibrium: They may 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 and carefully. 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 types do and do not exert effort. Notice that in the current case, opportunistic types have asymmetric incentive to exert effort when possessing different reputations. In particular, low-reputation opportunistic types have more incentive to exert effort, since they are more likely to get early moderation.

We are more interested in the case where opportunistic types may 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 types exert no effort when possessing high reputations, whereas (some) opportunistic types exert effort when possessing low reputations. We assume a proportion w ($w \leq 1/2$) of opportunistic types has high reputations and a proportion $1 - w$ has low reputations, and the number of opportunistic types with each reputation is invariant over time. Under such a scenario, it must be the case that low-reputation opportunistic types exert effort with probability $\frac{w}{1-w}$ to make the number of opportunistic types in high reputation stable. The expected success rate can be formulated as

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

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

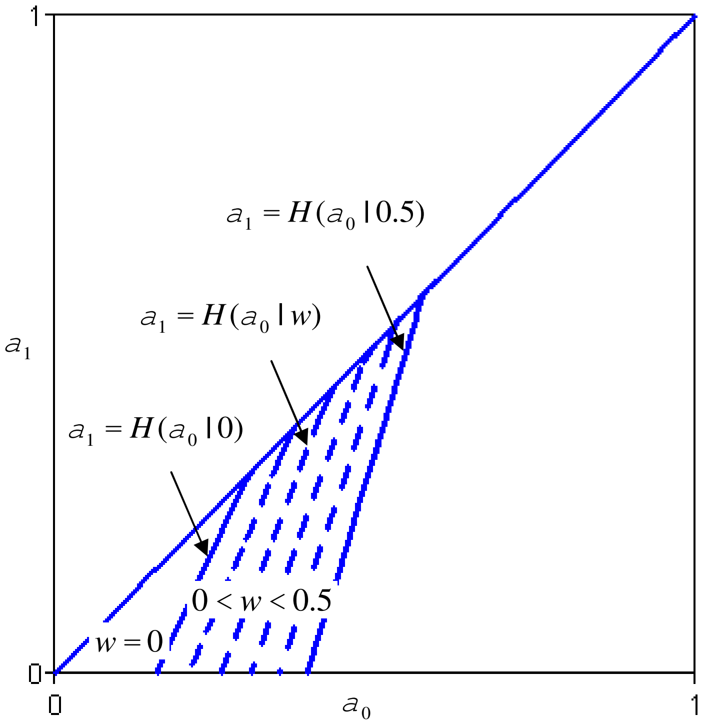

Figure 1: Equilibria under $\alpha_1 < \alpha_0$

We denote

$$H(\alpha_0|w) \equiv \frac{\beta[r_1(w) - (1-\alpha_0)\, r_0(w)] + \alpha_0 - c}{\beta r_1(w)}. \tag{17}$$

Such an equilibrium is characterized by the following proposition.

Proposition 3 (Reputation Oscillation) *For any (α_0, α_1) with $\alpha_1 < \alpha_0$, if $H(\alpha_0|0.5) \leq \alpha_1 \leq H(\alpha_0|0)$, the following is an equilibrium: In each period, a proportion w of opportunistic types are in high reputation and exert no effort, and the other opportunistic types are in low reputation and exert effort with probability $\frac{w}{1-w}$, where w is determined by $H(\alpha_0|w) = \alpha_1$.*

The proposition predicts that the reputations of opportunistic types oscillate between high and low: They build up high reputations when they are in low-reputation states, then exploit the reputation when they are in high-reputation states.

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

The condition $\alpha_1 \leq H(\alpha_0|0)$ is to make sure it is at least in some opportunistic types' interest to exert effort. In fact, when $\alpha_1 > H(\alpha_0|0)$, all opportunistic types stay at 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, $\alpha_1 = H(\alpha_0|0)$ defines the boundary condition beyond which no opportunistic types exert effort.

The condition $H(\alpha_0|0.5) \leq \alpha_1$ needed here is only to ensure that the number of opportunistic types in high reputation is stable. Without this condition, the reputation oscillation observed in equilibrium still holds. In fact, it is possible that in one period some opportunistic types have high reputations and exert no effort, and the rest of the opportunistic types have low reputations and all exert effort; in the next period, those two groups switch their roles, i.e., the high-reputation opportunistic types switch to low reputations and exert effort, while the low-reputation ones switch to high reputations and exert no effort. We take two groups of equal size as an example. Under such an equilibrium, the expected payoff for a high reputation is $v_1 = (1-\alpha_1)\, r_1(\frac{1}{2}) + \beta v_0$, and the expected payoff under a low reputation is $v_0 = (1-\alpha_0)\, r_0(\frac{1}{2}) + \alpha_0 - c + \beta v_1$. The incentive compatibility conditions $\alpha_1 + \beta(v_1 - v_0) - c < 0$ and $\alpha_0 + \beta(v_1 - v_0) - c > 0$ require

$$\alpha_1 < \min\left\{ H\left(\alpha_0|0.5\right), \frac{(1+\beta)\alpha_0 - \beta r_1(0.5) H(\alpha_0|0.5)}{1+\beta-\beta r_1(0.5)} \right\}. \tag{18}$$

The following example illustrates that condition (18) can be easily satisfied.

Example 1 *Let $\mu = s = 1/2$, $c = 1/4$, and $\beta = 4/5$. According to (15) and (16), we can calculate the expected performance $r_1(0.5) = 1/4$ and $r_0(0.5) = 3/4$. Suppose the moderation system does not moderate high-reputation commentators and moderates low-reputation commentators with probability $1/2$, i.e., $\alpha_1 = 0$ and $\alpha_0 = 1/2$. It is easy to verify that (18) holds, which means the equilibrium described above can be sustained as an equilibrium under the environment we specified.*

The above discussion shows the importance of moderation. In general, moderation plays a role in inducing opportunistic commentators' effort, and the frequency of moderation impacts opportunistic types' incentives to exert effort. As shown above, when low-reputation types are moderated more frequently, opportunistic types 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 ones. The equilibrium captured by (18) is clearly an example: r_0 is a weighted average of s and 1 by (16), whereas r_1 is a weighted average of s and 0 by (15), which implies $r_0 > r_1$. For the equilibrium in Proposition 3, we have the following result.

Corollary 2 (Reversed Reputation) *When the equilibrium w, determined by $H(\alpha_0|w) = \alpha_1$ in Proposition 3, is greater than $s/2$, the expected performance of high-reputation commentators is lower than that of low-reputation ones. Formally, $r_0(w) > r_1(w)$.*

In these scenarios, high reputation, in fact, means something "bad" (and in equilibrium readers anticipate that). This 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 impact 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, it may occur that, overall, the moderators check the low reputation more often. In such instances, readers should be informed of such a fact or be guided to read comments from low-reputation commentators first, since 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 an equilibrium with proportion w of the opportunistic type in high reputation and exerting effort,

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

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

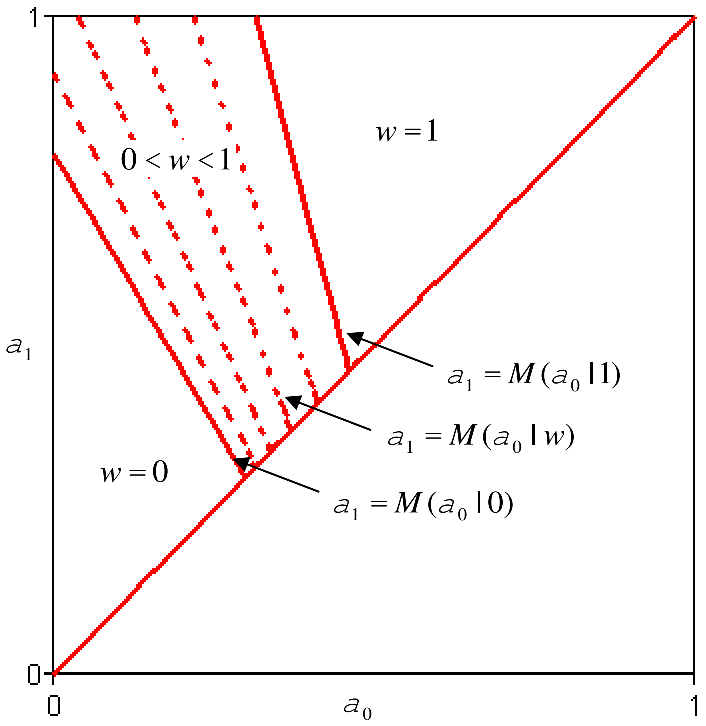

Figure 2: Equilibria under $\alpha_0 < \alpha_1$

The incentive conditions require that high-reputation opportunistic commentators are induced to exert effort while their low-reputation counterpart are not; formally,

$$\beta(v_1 - v_0) + \alpha_1 - c \geq 0 \text{ and } \beta(v_1 - v_0) + \alpha_0 - c \leq 0. \tag{21}$$

Unlike the case with more moderation on low reputation commentators, these incentive conditions are loose in the sense that multiple solutions may exist for a given (α_0, α_1); in other words, they may admit multiple equilibria. Intuitively, this is because a small portion of opportunistic commentators switching from one reputation to the other may not cause a change of the incentive structure. We next illustrate an equilibrium with binding incentive condition for low-reputation commentators, i.e., $\beta(v_1 - v_0) + \alpha_0 - c = 0$. By substituting $r_i(w)$ in, the incentive condition can be reorganized as

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

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 types in high reputations (see Figure 2).

When the moderation probability for high reputation is below a lower bound ($M(\alpha_0|0)$), no opportunistic commentators exert effort; and when the moderation is above a upper bound ($M(\alpha_0|1)$), all opportunistic commentators exert effort.

5 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 (8)), moderating comments with equal probability regardless of the commentators' reputations can induce opportunistic types 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 increased overall "system performance" and the cost of doing so. In other words, the community designer faces a decision on optimal moderation. We define the overall system performance as the average performance from each reputation weighted by its respective population size, $n_1r_1 + n_0r_0$, where n_i is the number of commentators with reputation i. Such a definition does measure the overall system performance, since 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 resource ($n_1\alpha_1 + n_0\alpha_0$) and denote it as $C(n_1\alpha_1 + n_0\alpha_0)$. Then the community designer's objective function can be formulated as

$$\max_{\alpha_1,\alpha_0}(n_1r_1 + n_0r_0) - C(n_1\alpha_1 + n_0\alpha_0) \tag{23}$$

We are interested in whether the moderation system should moderate high-reputation commentators more or low-reputation ones more. As mentioned earlier, for some given (α_1, α_0), multiple equilibria may exist, which could make it difficult to compare the system performance associated with each moderation scheme. To consider the optimal moderation problem, we here choose the minimum equilibrium system performance as the comparison criterion.

Proposition 4 *Considering the steady-state minimum-performance equilibrium, equally moderating all commentators is superior to moderating high-reputation commentators more, and moderating low reputation only is superior to any other moderation scheme that moderates low-reputation commentators more.*

When moderators pay more attention to high-reputation commentators, high-reputation opportunistic types have more incentives to exert effort than their low-reputation counterpart. In cases with some opportunistic types exerting effort, it must be that high-reputation opportunistic types exert effort and low-reputation opportunistic types not. In these cases, increasing moderation probability on low-reputation commentators (resulting in less freeriding) 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 high-reputation opportunistic types benefit from early moderation (by receiving maximum readership). Therefore, properly increasing moderation on low reputation and decreasing that on high reputation can keep opportunistic types' incentives unchanged and thus keep the system performance unchanged (recall the difference in reputation value influences commentators' incentives). In some sense, moderation on high reputation and that on low reputation are substitutes. This also explains why the slope of $\alpha_1 = M(\alpha_0|w)$ is negative in Figure 2. The above proposition shows the moderation on low reputation is a more effective one.

When moderators pay more attention to low-reputation commentators, high-reputation opportunistic types have less incentive to exert effort and may free ride. Unlike the previous case, reducing the moderation on high-reputation commentators here results in higher value for them due to the increased chance of freeriding. Meanwhile, reducing the moderation on low reputation can increase the value of staying at low reputation (consider the value from exerting no effort since they are indifferent). Therefore, properly reducing moderation on high reputation and reducing that on low reputation can keep opportunistic types' incentives unchanged and thus keep the system performance unchanged. So, the most cost effective approach is to moderate low reputation at the minimum probability for a certain level of performance. The lines $\alpha_1 = H(\alpha_0|w)$ in Figure 1 illustrate such intuition: lowering moderation probability on high reputations along these lines (to assume a same performance level) can lower moderation on low reputations; and thus the most cost effective way is to not moderate high reputations at all.

The above proposition predicts that it is optimal to either moderate high and low reputation with equal probabilities or moderate low reputation only. Numerical examples show that it may be optimal to moderate low-reputation commentators only, instead of moderating high and low reputation with equal probabilities.

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

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

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

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

Corollary 3 *When the total moderation α_T is exogeneously given and less than α_L, moderating low-reputation group only is optimal.*

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

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

It is worth noting that the system performance resulting from the optimal moderation discussed above is generally different from "efficient" system performance. The former is concerned about balancing the benefit and cost from the community designer's perspective, while the latter is from a social planner's perspective concerned about balancing the social value created (e.g., how much readers value the comments) and social cost incurred in generating those comments. Because of the hidden information regarding opportunistic types' effort level and costly moderation, the optimal system performance level is less than the efficient one in most cases.

6 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 impacts opportunistic types' 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.

6.1 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 moderators fairly judge a useful comment with probability p and always recognize worthless comments as worthless. Then the payoff functions in (3) can be re-formulated as[8]

$$v_i = \max_{e \in \{0,1\}} \left[\alpha ep + (1-\alpha) r_i\right] + \beta \left[epv_1 + (1-ep) v_0\right] - ce, \text{ for } i \in \{0,1\}. \tag{24}$$

Similar to (8) and (11), we can derive α_H and α_L under imperfect moderation as

$$\alpha_H = \frac{\frac{c}{p} - \beta\left(r_1 - r_0\right)}{1 - \beta\left(r_1 - r_0\right)}, \text{ and } \alpha_L = \frac{\frac{c}{p} - \beta\left(s - r_0\right)}{1 - \beta\left(s - r_0\right)}, \tag{25}$$

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

Proposition 5 *Both α_H and α_L decrease in p.*

Moderation provides opportunistic types more incentive to exert effort, since 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, since they are more likely to be fairly judged. Notice α_H measures the lower bound of moderation frequency to induce opportunistic types' 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.

The above proposition implies that it is beneficial for communities to improve the quality of moderation. Moderators could make mistakes due to their knowledge limitations or misunderstandings. In this sense, it is very critical to clearly state the community mission to commentators and especially to moderators. For example, what is the purpose of the online community? What kinds of comments (e.g., informative) are encouraged and what (e.g.,

[8] For simplicity, we continue to assume the verified worthless comments from early moderation get 0 readership, although some of those comments may be useful.

off-topic) are discouraged? Mis-moderation can also occur because moderators may have their own agenda, e.g., 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 post comments. To deal with this, Slashdot introduced a meta-moderation system, in which moderation may be judged as fair or unfair. Implementing another level of moderation, meaning moderation on moderators, is an effective way to insure the quality of moderation.

6.2 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 (3) can be re-formulated accordingly. For example,

$$v_0 = \max_{e\in\{0,1\}} \left[\theta\alpha e + (1-\theta\alpha)\, r_0\right] + \beta\left[\theta e v_1 + (1-\theta e)\, v_0\right] - ce. \tag{26}$$

We can conduct the same analysis to obtain similar results as those in our baseline model.

7 Conclusion

In this paper, we investigated the impact of a moderation system on the performance of online communities. First, we considered even moderation probability for different reputations and found that moderation probabilities critically impact 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. If a reputation system without moderation is viewed as a benchmark, we showed that the reputation system with moderation always outperforms the benchmark system. Then, we studied a model with differentiated moderation probabilities for different reputations, where we discovered reversed reputation and reputation oscillation. It was shown that agents in low-reputation category may exert more effort than those in high-reputation category, and then they exploit their reputations 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 only. We also illustrated that reputation oscillation and reversed 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 (Mayzlin, 2006). 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 better performance of the online community. Especially when moderation resources are limited (e.g., in terms of personnel, system capacity, and so on), resources should be directed toward users according to their reputation; it is generally effective to start with moderating one group of users with the same reputation.

Moderation has a potential to be applied to other fields. One direction is to cultivate an online moderated article review process. Online research networks, such as Social Science Research Network (SSRN), allow researchers to expose their working papers to the public. However, SSRN does not provide a platform for users to comment and discuss on a piece of work. The lack of such a functionality may be due to various reasons but one concern is the quality of comments. Our study on the moderation system can offer some guidelines for maintaining a quality research network, where researchers have incentive to provide their quality comments.

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 useful information. Such hidden abilities are due to various factors, such as their knowledge level and the opportunity cost of their effort/time needed. As a result, a simple reputation measure that only considers the recent moderation outcome is insufficient, since the reputation is not only 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, we can further consider that ownership of reputations can be shifted without indication. As we mentioned earlier, one important feature of online communities is the anonymity of users. Anonymity calls for reputation and/or moderation systems to regulate users' behavior. It also creates the more challenging issue of shift of reputations, which is beyond the scope of this study on reputations. After all, in online communities, an ID is just a symbol with some history, and we often do not know what happens behind these symbols.

Third, further empirical or experimental tests on our results remains to be considered as interesting direction. We are currently developing a system using PHP (PHP Hypertext Preprocessor), which is a cross-platform server-side programming language especially suited for web development. To maximize re-usability, the system is divided into two parts, a moderation system and a separate part containing contents. The moderation system is built subject independent, and it provides general moderation functions via API (Application Programming Interface) so that it can be applied to any environment.[9] The content part is built on top of the moderation system using the API. A completed system will provide a template for implementing a moderation system, and it will also allow us to collect first-hand data.

[9] Of course, many details need to be considered and are involved with information systems research. For example, we need to specify how the system calculates reputation scores based on users' past performance. Fan et al. (2005) provides us a good framework to start with.

A Appendix

A.1 Proof of Proposition 2

In the equilibrium with effort, it is easy to see $r_1 > r_0 = s$. Similarly, in the equilibrium with no effort, $r_1 = s > r_0$.

For the case where opportunistic types adopt a mixed strategy, notice the expected performance is the weighted average of the success rate s and m. The ratio of weight for s to m determines the value. We compare the weight ratios. If $m > s$, the expected performance is increasing in the ratio of weigh for m to that for s. So, to conclude $r_1 > r_0$, it is sufficient to show

$$\frac{(1-\mu)\, mp}{\mu sp} \geq \frac{(1-\mu)\,(1-mp)}{\mu\,(1-sp)} \tag{27}$$

or $\frac{m}{s} > \frac{1-mp}{1-sp}$. The latter is clearly true. If $m < s$, similarly, we can derive $r_1 > r_0$. $m = s$ is a special case in which $r_1 = r_0$.

A.2 Proof of Proposition 3

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

$$(\alpha_0 - c) + \beta\,[(1-\alpha_1)\, r_1(w) - (1-\alpha_0)\, r_0(w)] = 0. \tag{28}$$

Therefore,

$$\alpha_1 = \frac{\beta[r_1(w) - (1-\alpha_0)\, r_0(w)] + \alpha_0 - c}{\beta r_1(w)} = H(\alpha_0|w) = 1 - \frac{\beta\,(1-\alpha_0)\, r_0(w) - \alpha_0 + c}{\beta r_1(w)}. \tag{29}$$

Since $r_1(w)$ decreases in w and $r_0(w)$ increases in w, the right hand side is decreasing in w. So, if $H(\alpha_0|0.5) \leq \alpha_1 \leq H(\alpha_0|0)$, we can get a unique solution to $\alpha_1 = H(\alpha_0|w)$.

A.3 Proof of Corollary 2

By simple algebra, the condition for $r_1(w) < r_0(w)$ can reduce to $\mu s^2 < 2\mu sw + (1-\mu)\, w^2$, or $2\mu sw + (1-\mu)\, w^2 - \mu s^2 > 0$. Clearly, $2w > s$ is a sufficient condition.

A.4 Proof of Proposition 4

First, the system performance $(n_1 r_1 + n_0 r_0)$ is determined by the proportion of opportunistic types that exerts effort in equilibrium. In fact, if w is the proportion, $n_1 r_1 + n_0 r_0 = \mu s + (1 - \mu) w$. We next examine the minimum moderation resource required to achieve a proportion w, or

$$\min_{\alpha_1, \alpha_0} (n_1 \alpha_1 + n_0 \alpha_0) \tag{30}$$

For all (α_0, α_1) with $\alpha_1 \leq \alpha_0$, we consider the case with $w \leq 1/2$ here. For $w = 0$, the minimum resource required is trivially 0. For any $0 < w \leq 1/2$, notice that the equilibrium described in Proposition 3 simply implies that proportion w of opportunistic types exert effort in equilibrium. So, $\alpha_1 = H(\alpha_0|w)$ defines a line on which any pair (α_0, α_1) yields the same proportion w. Substituting $\alpha_1 = H(\alpha_0|w)$ into the above minimization problem, we can easily verify that the coefficient of α_0 is positive. Therefore, the optimal solution is minimum α_0 possible on $\alpha_1 = H(\alpha_0|w)$. Given the positive slope of $H(\alpha_0|w)$, the optimal solution is $\alpha_1 = 0$ and α_0 determined by $0 = H(\alpha_0|w)$. The other case with $w > 1/2$ can be similarly analyzed, and the optimal solution is $\alpha_1 = \alpha_0$.

For all (α_0, α_1) with $\alpha_1 \geq \alpha_0$, in an equilibrium with proportion w of the opportunistic type in high reputation and exerting effort, $v_1 = (1 - \alpha_1) r_1(w) + \alpha_1 - c + \beta v_1$ and $v_0 = (1 - \alpha_0) r_0(w) + \beta v_0$. Based on (19) and (20), by simple algebra we can show that either $(v_1 - v_0)$ decreases in w or $(v_1 - v_0)$ first increases and then decreases in w. Considering the minimum-performance equilibrium, we next show that a non-zero minimum proportion w_0 implies the low reputation incentive condition in (21) binds, i.e., $\beta(v_1 - v_0) + \alpha_0 - c = 0$. Suppose otherwise; that is $\beta(v_1 - v_0) + \alpha_0 - c < 0$ and $\beta(v_1 - v_0) + \alpha_1 - c \geq 0$. In the case that $(v_1 - v_0)$ decreases in w at $(w_0 - \epsilon, w_0)$, $w_0 - \epsilon$ can be supported as an equilibrium, and that is smaller than w_0, a contradiction. In the case $(v_1 - v_0)$ increases in w at $(w_0 - \epsilon, w_0)$, if $\beta(v_1 - v_0) + \alpha_1 - c > 0$, $w_0 - \epsilon$ can be supported as an equilibrium, and that is smaller than w_0, a contradiction; if $\beta(v_1 - v_0) + \alpha_1 - c = 0$, at $w = 0$ we have $\beta(v_1 - v_0) + \alpha_i - c < 0$, which means $w = 0$ can be sustained as an equilibrium, a contradiction. Therefore, we have $\beta(v_1 - v_0) + \alpha_0 - c = 0$, which leads to $\alpha_1 = M(\alpha_0|w)$ in (22). Then for $w = 0$, the minimum resource required is trivially 0. For any $0 < w \leq 1$, $\alpha_1 = M(\alpha_0|w)$ define a line on which any pair (α_0, α_1) yields the same proportion w. Substituting $\alpha_1 = M(\alpha_0|w)$ into the above minimization problem, clearly, the coefficient of α_0 is $\frac{-(1-\beta+\beta r_0(w))}{\beta(1-r_1(w))} n_1 + n_0$. We can verify this coefficient is negative, and thus the optimal solution is $\alpha_0 = \alpha_1$.

A.5 Proof of Proposition 5

We can re-organize α_H as follows,

$$\alpha_H = \frac{\frac{c}{p} - \beta\left(r_1 - r_0\right)}{1 - \beta\left(r_1 - r_0\right)} = 1 - \frac{1 - \frac{c}{p}}{1 - \beta\left(r_1 - r_0\right)}, \tag{31}$$

where

$$r_1 = \frac{\mu sps + (1-\mu)\,p}{\mu sp + (1-\mu)\,p}, \text{ and } r_0 = \frac{\mu\left(1 - sp\right)s + (1-\mu)\left(1-p\right)}{\mu\left(1-sp\right) + (1-\mu)\left(1-p\right)}. \tag{32}$$

First, $\frac{c}{p} < 1$ due to the assumption $c < p$. Notice that r_0 is the weighted average of the expected success rates s (from dedicated types) and 1 (from opportunistic types). So r_0 increases in $\frac{(1-\mu)(1-p)}{\mu(1-sp)}$, the weight ratio of 1 and s. It is easy to show that $\frac{(1-\mu)(1-p)}{\mu(1-sp)}$ decreases in p, so r_0 decreases in p. Also, notice that r_1 is independent of p. Then, for the second term on the right hand side of (31), the numerator increases in p and the denominator decreases in p, and thus it increases in p. So, a_H decreases in p.

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

References

Bateman, Patrick J., Peter H. Gray, Brian S. Butler. 2006. Community commitment: How affect, obligation, and necessity drive online behaviors. *Proceedings of Twenty-Seventh International Conference on Information Systems*. Milwaukee, Wisconsin, 983–1000.

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

Chromatic, Brian Aker, David Krieger. 2002. *Running Weblogs with Slash*. O'Reilly Media, Inc.

Cripps, Martin W., George J. Mailath, Larry Samuelson. 2004. Imperfect monitoring and impermanent reputations. *Econometrica* **72**(2) 407–432.

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

DeLone, William H., Ephraim R. McLean. 1992. Information systems success: The quest for the dependent variable. *Information Systems Research* **3**(1) 60–95.

Fan, Ming, Yong Tan, Andrew B. Whinston. 2005. Evaluation and design of online cooperative feedback mechanisms for reputation management. *IEEE Transactions on Knowledge and Data Engineering* **17**(2) 244–254.

Kreps, David M., Robert Wilson. 1982. Reputation and imperfect information. *Journal of Economic Theory* **27**(2) 253–279.

Lampe, Cliff, Paul Resnick. 2004. Slash(dot) and burn: Distributed moderation in a large online conversation space. *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems*. Vienna Austria, 543–550.

Mayzlin, Dina. 2006. Promotional chat on the internet. *Marketing Science* **25**(2) 155–163.

Milgrom, Paul, John Roberts. 1982. Predation, reputation, and entry deterrence. *Journal of Economic Theory* **27**(2) 280–312.

Parameswaran, Manoj, Andrew B. Whinston. 2007. Research issues in social computing. *Journal of AIS* **8**(6) 336–350.

Wasko, Molly McLure, Samer Faraj. 2005. Why should I share? examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly* **29**(1) 35–57.