Optimizing Discount & Reputation Trade-Offs in E-Commerce Systems: Characterization and Online Learning

Hong Xie,[1,2] **Yongkun Li,**[3] **John C.S. Lui**[2]
[1]College of Computer Science, Chongqing University, China
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[3]School of Computer Science and Technology, University of Science and Technology of China
hongx87@gmail.com, ykli@ustc.edu.cn, cslui@cse.cuhk.edu.hk

Abstract

Feedback-based reputation systems are widely deployed in E-commerce systems. Evidences showed that earning a reputable label (for sellers of such systems) may take a substantial amount of time and this implies a reduction of profit. We propose to enhance sellers' reputation via price discounts. However, the challenges are: *(1) The demands from buyers depend on both the discount and reputation; (2) The demands are unknown to the seller.* To address these challenges, we first formulate a profit maximization problem via a semi-Markov decision process (SMDP) to explore the optimal trade-offs in selecting price discounts. We prove the monotonicity of the optimal profit and optimal discount. Based on the monotonicity, we design a QLFP (Q-learning with forward projection) algorithm, which infers the optimal discount from historical transaction data. We conduct experiments on a dataset from to show that our QLFP algorithm improves the profit by as high as 50% over both the classical Q-learning and speedy Q-learning algorithm. Our QLFP algorithm also improves the profit by as high as four times over the case of not providing any price discount.

Introduction

Nowadays, E-commerce systems, e.g., Alibaba, Amazon, eBay and Taobao, are becoming increasingly popular. Such systems have generated tremendous economic values, e.g., Amazon and eBay were ranked 29th and 172nd in a Fortune 500 ranking (Fortune500 2015) in terms of the total revenue. This paper considers the eBay like E-commerce systems, where a large number of sellers and buyers transact online. To reflect the trustworthiness of sellers, a reputation system is maintained. In particular, the feedback-based reputation system (Resnick et al. 2000) is the most widely deployed, e.g., in eBay, Taobao, etc. Sellers of such systems are initialized with a low reputation and they must obtain a sufficiently large number of positive feedbacks from buyers to earn a reputable label. For example, eBay and Taobao use three-level feedbacks, i.e., $\{-1$ (Negative), 0 (Neutral), 1 (Positive) $\}$. Each seller is initialized with a reputation score of zero. A positive (or negative) rating increases (or decreases) the reputation score by one, while a neutral rating does not change the reputation score. To earn a 4-star label (i.e., a reputable label), a seller must increase her reputation score to at least 500 (eBay 1995).

Often, buyers are less willing to buy products from low reputation sellers. Authors in (Xie and Lui 2015) found that new sellers need to spend at least seven hundred days (on average) to earn a reputable label. Hence, some sellers resort to "*illegal means*" to increase their reputation, i.e., authors in (Xu et al. 2015) found that more than eleven thousand sellers in Tabobao have conducted fake transactions. A number of companies, e.g., Lantian, Shuake and Kusha, even provide professional fake transaction services and the per-year fake transaction volume is estimated to be more than six million per company (Xu et al. 2015). Fake transactions are illegal, and this motivates us to explore "*legitimate means*" to enhance (new) sellers' reputation.

We propose to enhance sellers' reputation via "*price discounts*". To illustrate, consider the eBay reputation system and that a seller is reputable if and only if her reputation score is no less than 500. A seller can attract 10 transactions per day if she is reputable, otherwise, she can only attract 1 transaction per day. Assume each transaction earns a positive rating of 1. Suppose the price of a product is \$1 and its cost is \$0.8. We have the following two cases:

Case 1 (No discounts) *For a new seller (initialized with a reputation score of zero) who does not provide any discount, she needs to spend 500 days to earn a reputable label. The total profit in the first 500 days is* $(1-0.8)\times 1\times 500 = 100$.

Case 2 (With discounts) *A new seller provides a discount of 40% before she earns a reputable label, i.e., the price becomes 0.6, and she does not provide any discount after becoming reputable. Assume this discount increases the transaction volume to 2 per day. She needs to spend* 250 *days to earn a reputable label. The profit in the first 250 days is* $(0.6-0.8)\times 2\times 250 = -100$. *The total profit in the first 500 days is* $(0.6-0.8)\times 2\times 250 + (1-0.8)\times 10\times 250 = 400$.

The above cases illustrate: (1) Price discounts can enhance sellers' reputation; (2) Price discounts may lead to profit losses in the short run, but the reputation effect can compensate the profit in subsequent days. Note that in real-world E-commerce systems, the demands (i.e., per-day transaction volumes) are *dynamic*, buyers may provide *biased* ratings, and the discount-dependent demands (i.e., buyers' preferences over discounts) are *unknown*, etc. This paper studies the discount selection problem in such general settings, and we aim to answer: *(1) What are the optimal trade-offs in selecting price discounts? (2) How to perform online inference to determine the optimal discount from historical transaction data?* Our contributions are:

- We develop a mathematical model to capture important factors of an E-commerce system, i.e., the *demand dynamics*, *rating biases*, and *buyers' preferences* over discounts, etc. We formulate a profit maximization framework via an SMDP to explore the optimal trade-offs in determining the optimal price discount.
- We prove the monotonicity of the optimal profit and discount via *convex optimization* (Boyd and Vandenberghe 2004) and *comparative statics* (Chiang 1984). Based on the monotonicity, we design a QLFP (Q-learning with forward projection) algorithm, which infers the optimal discount from historical transaction data.
- We conduct experiments on a dataset from eBay to show that our QLFP algorithm improves the profit by as high as 50% over both the classical Q-learning and the speedy Q-learning algorithm. Our QLFP algorithm also improves the profit by as high as four times over the case of not providing any price discount.

This remaining of this paper organizes as follows. We first present the system model and the problem formulation. Then, we theoretically characterize the optimal profit and price discount. Based on these characterizations, we design of our QLFP algorithm. We conduct experiments on a data set from eBay to evaluate our QLFP algorithm. Finally, we present the related work and conclusion.

System Model

We first model the baseline E-commerce system and buyers' preferences over price discounts. We then formulate a profit maximization framework via an SMDP to characterize the optimal trade-offs in selecting price discounts. Finally, we formulate an online discount selection problem, which infers the optimal discount from a seller's transaction data, without knowing any buyer's preference over discounts.

Baseline E-commerce System Model

Consider an E-commerce system like eBay and Taobao where buyers purchase products from online stores operated by sellers, and a feedback-based reputation system is maintained to reflect the trustworthiness of sellers. Sellers set the selling price and advertise the quality of products in their online stores, and finally ship the ordered products to buyers. Let $q \in \mathbb{R}_+$ and $c \in \mathbb{R}_+$ denote the price and overall cost of a product respectively. The overall cost c captures the manufacturing cost, shipment fee, etc. We define the *unit profit* to the seller $u \in \mathbb{R}$ as the price minus the cost, i.e., $u \triangleq q - c$. Sellers advertise product quality honestly and we aim to enhance sellers' reputation via price discounts.

To reflect the trustworthiness of sellers, a feedback-based reputation system tags each seller with a reputation score $s \in \mathcal{S}$, and this score is accessible by all buyers, where

$$\mathcal{S} \triangleq \left\{-\hat{S}, \ldots, -1, 0, 1, \ldots, S\right\},$$

and $\hat{S}, S \in \mathbb{N} \cup \{\infty\}$. For example, eBay and Taobao uses $\hat{S} = 0, S = \infty$, in other words $\mathcal{S} = \{0, 1, \ldots, \infty\}$. The higher the reputation score, the more reputable the seller is. When a seller joins an E-commerce system, the reputation system initializes her reputation score as $s = 0$, i.e., a low reputation. Buyers provide feedback ratings to reflect their evaluation on the overall transaction quality (i.e., product quality, trustworthiness of the seller, etc). Each feedback rating is drawn from a discrete rating metric set

$$\mathcal{M} \triangleq \left\{-\hat{M}, \ldots, -1, 0, 1, \ldots, M\right\},$$

where $\hat{M}, M \in \mathbb{N}$. For example, eBay and Taobao deploy $\mathcal{M} = \{-1(\text{Negative}), 0(\text{Neutral}), 1(\text{Positive})\}$. The higher the rating, the more satisfied the buyer is toward that seller. Consider a seller has a reputation score s, her reputation score becomes $s + m$ once she receives a feedback rating $m \in \mathcal{M}$. For example, in eBay $\mathcal{M} = \{-1, 0, 1\}$, a rating of 1 (or -1) increases (or decreases) the reputation score by 1, while a rating of 0 does not change the reputation score.

Price Discount Model

To speed up the reputation accumulating process, a seller can set a price discount $a \in \mathcal{A} \triangleq [0, 1]$. Precisely, a denotes the discount rate, and the product price under discount a is $q \times (1 - a)$. For example, $a = 0.2$ means 20% off and the corresponding price is $0.8q$. Also $a = 0$ captures that a seller does *not* provide any discount. Let $\tilde{u}(a)$ denote the *unit profit* under discount a. Then we have

$$\tilde{u}(a) \triangleq u - aq, \qquad \forall a \in \mathcal{A}.$$

Modeling rating behavior under discounts. Human factors like personal preferences or even biases need to be included in our model. Some buyers may provide higher ratings while other may provide lower ratings. Let $R(s, a) \in \mathcal{M}$ denote a rating provided by buyers to the seller who has a reputation score $s \in \mathcal{S}$ and sets a discount $a \in \mathcal{A}$. The rating $R(s, a)$ is a random variable, and we define its cumulative distribution function (CDF) as

$$F_R(m|s, a) \triangleq \mathbb{P}\left[R(s, a) \leq m\right], \quad \forall m \in \mathcal{M}, s \in \mathcal{S}, a \in \mathcal{A}.$$

Note that sellers do *not have any a-priori knowledge* on $F_R(m|s, a)$. For example, consider $\mathcal{M} = \{-1, 0, 1\}$ and $\mathcal{S} = \{0, 1, \ldots, \infty\}$. Then, one example of $F_R(m|s, a)$ is

$$\begin{cases} F_R(-1|s, a) = [0.1/(1 + s)]^{1+a}, \\ F_R(0|s, a) = [0.3/(1 + s)]^{1+a}, \\ F_R(1|s, a) = 1. \end{cases} \tag{1}$$

Definition 1 *Given two random variables X, Y with the same sample space Ω. We say X is larger than Y (written as $X \succeq Y$), iff $\mathbb{P}[X > x] \geq \mathbb{P}[Y > x]$ holds for all $x \in \Omega$.*

Assumption 1 *Given $a \in \mathcal{A}$, $R(s, a) \succeq R(j, a)$ holds for all $s > j$, where $s, j \in \mathcal{S}$. Given $s \in \mathcal{S}$, $R(s, a) \succeq R(s, b)$ holds for all $a > b$, where $a, b \in \mathcal{A}$.*

Assumption 1 captures: (1) The herding behavior (Muchnik, Aral, and Taylor 2013) that buyers give higher ratings to more reputable sellers; (2) The price effect that buyers tend to become more lenient in providing ratings under larger discounts. Equation (1) satisfies Assumption 1.

Modeling demand under discounts. We consider a dynamic demand from buyers and use the transaction's arrival process to model the demand. We define the transaction's arrival process through the inter-arrival time (or waiting time) of transactions. Precisely, let $W(s,a) \in \mathbb{R}_+$ denote the inter-arrival time of transactions to the seller who has a reputation score $s \in \mathcal{S}$ and sets a discount $a \in \mathcal{A}$. For example, $W(0,0)$ measures the amount of time a seller must wait until the next transaction arrives when she has a reputation score $s = 0$ and does not provide any discount. The inter-arrival time $W(s,a)$ is a random variable and we denote its CDF as

$$F_W(w|s,a) \triangleq \mathbb{P}[W(s,a) \le w], \quad \forall w \in \mathbb{R}_+, s \in \mathcal{S}, a \in \mathcal{A}.$$

Note that sellers do *not have any a-priori knowledge* on $F_W(w|s,a)$. One example of $F_W(w|s,a)$ is

$$F_W(w|s,a) = 1 - \exp(-\lambda(s,a)w), \tag{2}$$

which means that $W(s,a)$ follows an exponential distribution with a parameter $\lambda(s,a) \in \mathbb{R}_+$. This also models the Poisson arrival of transactions. One example of $\lambda(s,a)$ is

$$\lambda(s,a) = \frac{1+\sqrt{a}}{1+e^{-s}}, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{3}$$

Assumption 2 *Given $a \in \mathcal{A}$, $W(j,a) \succeq W(s,a)$ holds for all $s > j$, where $s,j \in \mathcal{S}$. Given $s \in \mathcal{S}$, $W(s,b) \succeq W(s,a)$ holds for all $a > b$, where $a,b \in \mathcal{A}$.*

Assumption 2 captures: (1) The reputation effect that buyers are more willing to transact with reputable sellers; (2) The price effect that buyers are more willing to buy a product under a larger discount. Consider Eq. (2), Assumption 2 means that $\lambda(s,a)$ increases in both s and a. One example of such $\lambda(s,a)$ is derived in Eq. (3).

Assumption 3 *There exists two constants $\epsilon > 0, \delta > 0$ such that $F_W(\delta|s,a) \le 1-\epsilon$ for all $s \in \mathcal{S}, a \in \mathcal{A}$.*

Assumption 3 states that it is impossible that an infinite number of transactions arrive to an online store within a finite time. Consider Eq. (2), Assumption 3 means that $\lambda(s,a)$ is bounded, e.g., the $\lambda(s,a)$ derived in Eq. (3).

Modeling discount update. This paper aims to enhance sellers' reputation via price discounts. The challenge is that sellers do *not have any a-priori knowledge* on $F_R(m|s,a)$ and $F_W(w|s,a)$. However, a seller can infer them from historical transaction data to optimize the price discounts. We therefore focus on the scenario that a seller updates the discount only when a new transaction arrives, i.e., gains some new data for inference. Under this scenario, we next introduce the formal discount selection models for sellers.

The Seller's Decision Model

The seller needs to select a discount for each transaction. Thus the decision space for the seller is the discount set $\mathcal{A}$.

Offline decision model. We first consider the full information scenario that $F_R(m|s,a)$ and $F_W(w|s,a)$ are given. We formulate a profit maximization framework via an SMDP to characterize the optimal trade-offs in determining discounts.

We consider a continuous time system with infinite-horizon $t \in [0,\infty)$. Let t_i denote the arrival time of the i-th transaction, where $i \in \mathbb{N}_+$. We say a seller is at state $s \in \mathcal{S}$ if she has a reputation score s. Thus, the state space is $\mathcal{S}$. Decision epochs correspond to the time immediately following an arrival of a transaction. For example, the first decision epoch occurs at t_1. The initial decision epoch does not correspond to any transaction. Without any loss of generality, we index the initial decision epoch with 0, and use $t_0 = 0$ to denote its occurrence time. The seller is the decision maker and the decision to be made at each decision epoch is setting a discount $a \in \mathcal{A}$. We also call a the action. Note that the action set at each decision epoch is the same $\mathcal{A}$. When the seller chooses an action a at state s, she receives a lump sum reward denoted by $k(s,a)$, which can be expressed as

$$k(s,a) = \tilde{u}(a), \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

Note that the lump sum reward corresponds to the unit profit earned from the next transaction. Namely, it is delayed to be payed in the next decision epoch.

Note that the inter-arrival (or waiting) time of decision epochs is $W(s,a)$, which is a random variable variable and has a CDF $F_W(w|s,a)$. Let $p(j|s,a)$, where $s,j \in \mathcal{S}, a \in \mathcal{A}$, denote the state transition probability

$$\begin{aligned} p(j|s,a) &\triangleq \mathbb{P}[\text{next state is } j|\text{current state } s, \text{ discount } a] \\ &= F_R(j-s|s,a) - F_R(j-s-1|s,a). \end{aligned}$$

Namely, $p(j|s,a)$ models the dynamics of the reputation score.

Setting price discounts may lead to some profit losses at the present decision epoch, but it can speed up the reputation score accumulation process, which may improve sellers' profit in subsequent decision epochs. To quantify the optimal discount and reputation trade-off, we use an *expected infinite-horizon discounted profit* for the seller. Precisely, we consider a continuous-time discounting rate $\alpha \in \mathbb{R}_+$ and define the expected infinite-horizon discounted profit as

$$v^\pi(s) \triangleq \mathbb{E}\left[\sum_{i=0}^{\infty} e^{-\alpha t_{i+1}} k(s_i,a_i) \middle| s_0 = s, \pi\right], \quad \forall s \in \mathcal{S},$$

where s_i, a_i denote the reputation score and discount at the i-th decision epoch, and π denotes a policy (Puterman 2014), which prescribes a discount for each transaction (or decision epoch). We also call $v^\pi(s)$ the *long-term profit*. For example, the long term profit for a new seller is $v^\pi(0)$. One interpretation of the discounting rate α is inflation from economic perspectives. The discounting rate α also reflects the willingness of a seller to trade discounts for reputation. Increasing α means that the seller cares less about the future profit (or is more keen about the present profit). In other words, she is less willing to trade discounts for reputation.

We define a stationary and deterministic (SD) policy as $\pi = (d)^\infty$, where $d : \mathcal{S} \to \mathcal{A}$ denotes a Markovian deterministic decision rule, which maps each state to a price discount.

Problem 1 (Offline discount selection) *Given the initial state s_0, $F_R(m|s,a)$, and $F_W(w|s,a)$, select price discounts to maximize the long term profit. Formally,*

$$\begin{aligned}&\underset{\pi}{\textit{maximize}} && v^{\pi}(s_0)\\ &\textit{subject to} && \pi \in \Pi,\end{aligned}$$

where Π denotes a set of all possible SD policies.

Problem 1 optimizes the long term profit over a special class of policies, i.e., SD policies, because SD policies suffice to attain the global maximum long term profit.

Online decision model. Now we relax problem 1 to the online decision making setting, in which $F_W(w|s,a)$ and $F_R(m|s,a)$ of problem 1 are *unknown* to the seller. The seller can only access her own historical transaction data and use her data to predict the optimal discount. Precisely, the i-the transaction data item is associated with: (1) A discount a_{i-1}, (note that the discount of a transaction is set in the last decision epoch); (2) A reputation score s_{i-1} at which the seller sets a_{i-1}; (3) A lump sum reward (i.e., profit) $k(s_{i-1}, a_{i-1})$; (4) An arrival time t_i; (5) A rating denoted by m_i. For example, at the 0-th decision epoch (i.e., the initial decision epoch), the seller sets a discount a_0 at state s_0. When the first transaction occurs at time t_1, the seller obtains a lump sum reward (i.e., profit) $k(s_0, a_0)$ and receives a rating m_1. The first transaction data item is then $\mathcal{H}_1 \triangleq \{a_0, s_0, k(s_0, a_0), t_1, m_1\}$. In general, the i-th transaction data item is $\mathcal{H}_i \triangleq \{a_{i-1}, s_{i-1}, k(s_{i-1}, a_{i-1}), t_i, m_i\}$, $i = 1, 2, \ldots, \infty$. For the ease of presentation, we define $\mathcal{H}_0 \triangleq \{t_0, s_0\}$ for the initial decision epoch. At the i-th decision epoch, a seller observes $\mathcal{H}_i$ and she uses it to infer the optimal discount.

Problem 2 (Online discount selection) *Given an initial state s_0, at the i-th decision epoch, where $i = 0, 1, \ldots, \infty$,*

- *receive $\mathcal{H}_i$ and determine a discount a_i based on $\mathcal{H}_i$,*

to maximize long term profit $\mathbb{E}\left[\sum_{i=0}^{\infty} e^{-\alpha t_{i+1}} k(s_i, a_i) | s_0\right]$.

We will first study Problem 1. Through this we lay the foundation to address Problem 2.

Optimal Profit and Discounts

It is mathematically intractable to derive the closed-form expression for the maximum long-term profit denoted by $v^*(s)$. In the following theorem, we identify a monotone property of $v^*(s)$.

Theorem 1 *For all $s \geq j$, where $s, j \in \mathcal{S}$, $v^*(s) \geq v^*(j)$ holds. Furthermore, $v^*(s)$ is non-increasing in α.*

Theorem 1 states that the seller can earn more profit if her reputation score increases or the inflation decreases. In other words, sellers always have incentive to increase their reputation scores. These monotone properties serve as an important building block for us to characterize the optimal discount. *Due to page limit, we present all proofs in our technical report (Xie, Li, and Lui 2018).*

Definition 2 *For each reputation score $s \in \mathcal{S}$, we define the associated action-dependent long term profit as $Q(s,a) \triangleq \phi(s,a)V(s,a)$, where $\phi(s,a) = \int_0^{\infty} e^{-\alpha w} dF_W(w|s,a)$ and $V(s,a) = k(s,a) + \sum_{j \in \mathcal{S}} p(j|s,a)v^*(j)$.*

Given that a seller has a reputation score s, the $Q(s,a)$ gives the maximum long term profit she can earn by setting a discount a. The optimal discount $d^*(s)$ satisfies $d^*(s) \in \arg\max_{a \in \mathcal{A}} Q(s,a)$.

Theorem 2 *Suppose $a \in \mathcal{A}_s$, where $\mathcal{A}_s$ is defined as $\mathcal{A}_s \triangleq \{a | Q(s,a) \geq 0, a \in \mathcal{A}\}$. For all $j > \ell \geq s$, where $j, \ell, s \in \mathcal{S}$, $Q(j,a) \geq Q(\ell,a)$ holds.*

Theorem 2 states that given the same discount $a \in \mathcal{A}_s$, the seller can earn more profit if she has a higher reputation score. We formulate the following problem to further study the optimal discount.

Problem 3 *Given s, select a to maximize $\ln Q(s,a)$:*

$$\textit{maximize}_{a \in \mathcal{A}} \quad \ln Q(s,a) = \ln \phi(s,a) + \ln V(s,a)$$

In Problem 3, we maximize the log function of the action-dependent long term profit. This treatment does not change the optimal discount and will facilitate the analysis.

Theorem 3 *Suppose $F_W(w|s,a)$ is strictly concave with respect to a and $F_R(m|s,a)$ is convex with respect to a. Problem 3 has a unique optimal solution.*

Theorem 3 derives sufficient conditions to guarantee the uniqueness of the optimal discount for each given s. This uniqueness enables us to further characterize the optimal discount via *comparative statics*. When the optimal discount is unique, it is algorithmically easy to locate it. For example, Eq. (1) satisfies the condition on $F_R(m|s,a)$.

Corollary 1 *Suppose $F_W(w|s,a)$ satisfies Eq. (2). If $\lambda(s,a)$ is strictly concave in a and $F_R(m|s,a)$ is convex in a, there exist a unique optimal discount for s.*

Corollary 1 states that given the Poisson arrival of transactions, if the transaction's arrival rate $\lambda(s,a)$ has a diminishing return in the discount a, then the optimal discount is unique for each reputation score. For example, Eq. (3) satisfies the condition on $\lambda(s,a)$.

In order to apply comparative statics to further characterize the optimal discount, we define the following notation.

Definition 3 *We define the hazard function of $Q(s,a)$ with respect to a as*

$$h(s,a) \triangleq -\frac{\partial Q(s,a)}{\partial a}\frac{1}{Q(s,a)}, \forall s \in \mathcal{S}, a \in \mathcal{A}_s.$$

The hazard function $h(s,a)$ measures the proportional reduction in the discount-dependent long-term profit (i.e., $-\partial Q(s,a)/Q(s,a)$) with respect to the marginal change in the price discounts (i.e., ∂a).

Theorem 4 *Suppose the conditions in Theorem 3 hold. If $h(s,a)$ is non-decreasing in α, the unique optimal discount $d^*(s)$ is non-increasing in α. If $h(s,a)$ is non-decreasing in s, the unique optimal discount $d^*(s)$ is non-increasing in s.*

Theorem 4 states sufficient conditions under which the unique discount is non-increasing in the discounting rate α and non-increasing in reputation score s. One interpretation is that the seller sets smaller discounts when the inflation increases or she is more keen about the present profit. More reputable sellers set smaller discounts.

Online Discount Selection

We first apply the Q-learning algorithm to infer the optimal discount from historical transaction data. To speed up the convergence, we design a QLFP algorithm, which extends the Q-learning to incorporate the characterizations in the last section.

Algorithm 1 : Discount Selection Via Q-learning

```
Require: Discounting rate $\alpha$, learning rate $\eta_i$, exploration
    probability $\epsilon_i$, initialization $Q^{(0)}(s,a)$;
1: for $i = 1$ to $\infty$ do
2:     Compute the waiting time $w_i \leftarrow t_i - t_{i-1}$.
3:     $\hat{\phi}(s_{i-1}, a_{i-1}) \leftarrow e^{-\alpha w_i}$.
4:     $\hat{r}(s_{i-1}, a_{i-1}) \leftarrow e^{-\alpha w_i}(u - a_{i-1}q)$.
5:     Update reputation score $s_i \leftarrow s_{i-1} + m_i$.
       If $s_i < -\hat{S}$, $s_i \leftarrow -\hat{S}$. If $s_i > S$, $s_i \leftarrow S$.
6:     $Q^{(i)}(s_{i1}, a_{i-1}) \leftarrow \hat{\phi}(s_{i-1}, a_{i-1}) \max_{a \in \mathcal{A}} Q^{(i-1)}(s_i, a) + \hat{r}(s_{i-1}, a_{i-1})$.
7:     If $s \neq s_{i-1}$ or $a \neq a_{i-1}$, $Q^{(i)}(s,a) \leftarrow Q^{(i-1)}(s,a)$, otherwise $Q^{(i)}(s_{i-1}, a_{i-1}) \leftarrow \eta_{i-1} Q^{(i)}(s_{i-1}, a_{i-1}) + (1-\eta_{i-1}) Q^{(i-1)}(s_{i-1}, a_{i-1})$.
8:     With probability $\epsilon_i$, $a_i \sim \text{UniformRandom}(\mathcal{A})$, with
       probability $1 - \epsilon_i$, $a_i \in \arg\max_{a \in \mathcal{A}} Q^{(i)}(s_i, a)$.
9: end for
```

Q-learning for online discount selection. We apply the Q-learning algorithm (Bradtke and Duff 1994) to infer the optimal discount. Recall that for each given reputation score s, the optimal discount $d^*(s)$ maximizes the $Q(s,a)$ and that in each decision epoch the seller observes the transaction data $\mathcal{H}_i$. Once receives $\mathcal{H}_i$, the seller first uses it to estimate $Q(s,a)$, and then selects a discount based on the estimated $Q(s,a)$. We formally outline this idea in Algorithm 1. To illustrate, suppose a seller is in the i-th decision epoch, i.e., receives $\mathcal{H}_i \triangleq \{a_{i-1}, s_{i-1}, k(s_{i-1}, a_{i-1}), t_i, m_i\}$. Step 2 computes the waiting time of the i-th transaction $w_i = t_i - t_{i-1}$. Step 3 estimates the per-epoch discount factor, i.e., $\hat{\phi}(s_{i-1}, a_{i-1}) = e^{-\alpha w_i}$. Step 4 estimates the per-epoch discounted profit, i.e., $\hat{r}(s_{i-1}, a_{i-1}) = \hat{\phi}(s_{i-1}, a_{i-1})k(s_{i-1}, a_{i-1}) = e^{-\alpha w_i}(u - a_{i-1}q)$. Step 5 updates the reputation score. Step 6 computes a new estimation of the $Q(s_{i-1}, a_{i-1})$ as $Q^{(i)}(s_{i-1}, a_{i-1}) = \hat{r}(s_{i-1}, a_{i-1}) + \hat{\phi}(s_{i-1}, a_{i-1}) \max_{a \in \mathcal{A}} Q^{(i-1)}(s_i, a)$. Step 7 updates the estimation of $Q(s_{i-1}, a_{i-1})$ by combining the old $Q^{(i-1)}(s_{i-1}, a_{i-1})$ and new estimation $Q^{(i)}(s_{i-1}, a_{i-1})$ with a learning rate $\eta_i \in \mathbb{R}_+$. Step 8 selects a discount to maximize $Q^{(i)}(s_i, a)$ with probability $1 - \epsilon_i$, and it selects a discount uniformly at random with probability ϵ_i (i.e., this corresponds to the *exploration* step in reinforcement learning). Note that Algorithm 1 is suitable for finite discount set $\mathcal{A}$ and finite reputation score set $\mathcal{S}$, because we need to store $Q(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. For our problem, we can discretize the discount set and truncate the reputation score to be finite. Under mild assumptions on rating bias $F_R(m|s,a)$ and proper selections of the exploration parameter ϵ_i and learning rate η_i, Algorithm 1 converges to the optimal policy, i.e., it selects the optimal discount asymptotically. Due to page limit, we present the convergence analysis in out technical report (Xie, Li, and Lui 2018).

Q-learning with Forward Projection (QLFP). Improving the above Q-learning algorithm, i.e., Algorithm 1, can improve a seller's profit. We now apply the insights obtained in last section to improve Algorithm 1. Recall that Theorem 2 states that given a reputation score s and a discount a, if $Q(s,a) \geq 0$, $Q(j+1, a) \geq Q(j, a)$ holdes for all $j \geq s$. Algorithm 2 applies this observation to further improve the prediction of $Q(s,a)$ via forward projection, which we call QLFP for short. For the input of Algorithm 2, we require the initial $Q^{(0)}(s,a)$ to satisfy Theorem 2. Step 2 executes the steps 2–7 of Algorithm 1 to obtain an estimation of $Q^{(i)}(s,a)$ based on $\mathcal{H}_i$. Step 3-7 makes the $Q^{(i)}(s,a)$ to satisfy Theorem 2 via *forward projection*, i.e., propagate the value of $Q^{(i)}(s_{i-1}, a_{i-1})$ upward in terms of the reputation score.

Algorithm 2 : QLFP Algorithm

```
Require: $\alpha$, $\eta_i$, $\epsilon_i$, $Q^{(0)}(s,a)$ (satisfies Theorem 2);
1: for $i = 0, 1$ to $\infty$ do
2:     Execute step 2–7 of Algorithm 1.
3:     if $Q^{(i)}(s_{i-1}, a_{i-1}) \geq 0$ then
4:         for $j = s_{i-1} + 1$ to $S$ do
5:             If $Q^{(i)}(j, a_{i-1}) < Q^{(i)}(j-1, a_{i-1})$,
               $Q^{(i)}(j, a_{i-1}) \leftarrow Q^{(i)}(j-1, a_{i-1})$.
6:         end for
7:     end if
8:     Execute step 8 of Algorithm 1.
9: end for
```

Under mild assumptions on rating bias $F_R(m|s,a)$ and proper selections of the exploration parameter ϵ_i and learning rate η_i, Algorithm 2 converges to an optimal policy. Due to page limit, we present the convergence analysis in our technical report (Xie, Li, and Lui 2018) We next conduct experiments to evaluate the convergence speed of our QLFP algorithm as well as its effectiveness in optimizing the reputation and discount trade-offs.

Experiments on Real Data

We conduct experiments on a dataset from eBay and show that our QLFP improve the profit by as high as 50% over Q-learning and Speedy Q-learning, and by as high as four times over the case of not providing any price discount.

Experiment Settings

Datasets. We use a dataset from eBay (Xie and Lui 2017), which contains 19,217,083 transactions of 4,586 sellers. For each seller, the dataset contains all her transactions up to April 2013. Each transaction data item contains a sellerID, a buyerID, a time stamp and a feedback rating provided by buyers. Each feedback rating is drawn from $\{-1, 0, 1\}$. Figure 1 plots the distribution of the number of transactions.

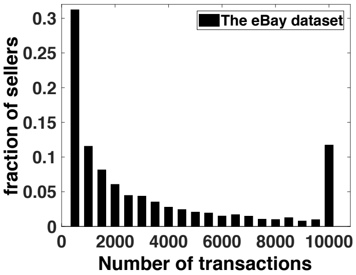


Figure 1: The distribution of number of transactions.

Model parameters. To assist buyers to assess sellers' reputation, eBay adopts a twelve-star label system (eBay 1995) summarized in Table 1. Authors in (Xie and Lui 2017) found that the transactions in eBay follow a Poisson arrival process. Thus, we consider a Poisson arrival of transactions and we infer the transaction's rate (without discounts) across the number stars via the empirical mean

$$\text{Trans. rate}|_{n\text{ stars}} = \frac{\#\text{ of trans. to sellers with } n \text{ stars}}{\text{total time to accumulate these trans.}}.$$

Table 2 presents the inferred per-day transactions' rate. From Table 2, one can observe that when the number of stars is less than 4, the transaction's rate is less than one per day. This verifies that when the reputation is low, it is difficult for sellers to attract buyers. Note that no seller has ever achieved a reputation score of more than 500,000, i.e., the number of 11 or 12 stars. Thus, the transaction's rate for these stars are missing. We synthesize the corresponding transaction's rate to capture that further increasing the reputation of highly reputable sellers increases the transactions slightly, i.e.,

# stars	0	1	2	3	4	5	6	7	8	9	10	11	12
min#rat	0	10	50	100	500	10^3	$5{\cdot}10^3$	10^4	$2.5{\cdot}10^4$	$5{\cdot}10^4$	10^5	$5{\cdot}10^5$	10^6

Table 1: Reputation score vs. the number of stars.

# stars	0	1	2	3	4	5	6	7	8	9	10	11	12
tran rate	0.05	0.18	0.33	0.68	1.29	2.37	4.57	8.13	15.59	28.69	89.39	–	–

Table 2: Transaction's rate across number of stars.

$$\text{Trans. rate}|_{11\text{ stars}} = 1.1 \times \text{Trans. rate}|_{10\text{ stars}} = 98.329$$
$$\text{Trans. rate}|_{12\text{ stars}} = 1.05 \times \text{Trans. rate}|_{11\text{ stars}} = 103.245$$

In eBay, sellers with reputation score 10^6 or above have the same number of stars, i.e., twelve stars. We thus truncate the reputation score set to be $\mathcal{S} = \{0, 1, \ldots, 10^6\}$. Let $\tilde{\lambda}_s$ denote the transaction's rate to a seller who has a reputation score s and does not provide any discounts. We infer it as the empirical transaction's rate, i.e.,

$$\tilde{\lambda}_s = \text{Trans. rate}|_{n\text{ stars}}, \quad \forall s \text{ is associated with } n \text{ stars}.$$

Note that $\mathcal{M} = \{-1, 0, 1\}$ for eBay. The fraction of each rating level in our dataset can be summarized as follows: 0.23% are of -1, 0.34% are of 0, and 99.43% are of 1. This implies a very small bias in providing feedback ratings. Thus, we set the rating distribution as

$$F_R(-1|s,a) = 0.0023, \\ F_R(0|s,a) = 0.0057, F_R(1|s,a) = 1, \qquad (4)$$

holds for all $s \in \mathcal{S}, a \in \mathcal{A}$.

To study the impact of rating bias in general, we also synthesize the feedback rating as

$$\begin{cases} F_R(-1|s,m) = \left[1/(1+\eta+\eta^2)\right]^{1+\gamma a}, \\ F_R(0|s,a) = \left[(1+\eta)/(1+\eta+\eta^2)\right]^{1+\gamma a}, \\ F_R(1|s,a) = 1, \end{cases} \qquad (5)$$

where $\eta = \theta + \ln(1+s)$, $\theta \in [1, \infty)$ and $\gamma \in \mathbb{R}_+$. For example, when $s = 0$, $a = 0$ and $\theta = 1$, the rating will be of $-1, 0, 1$ with equal probability 1/3. The θ models the baseline rating bias under no discounts. The larger the θ, the higher the probability of providing a high rating, i.e., a smaller rating bias. The γ models the sensitivity of buyers' rating leniency over discounts. The larger the γ, the higher the probability of providing a high rating.

We normalize the baseline price to be $q = 1$. Furthermore, we set the cost and the discount set to be $c = 0.6, \mathcal{A} = \{0.02k | k = 0, 1, \ldots, 25\}$. With discounts, we still consider a Poisson arrival of transactions, i.e., $F_W(w|s,a)$ satisfies Eq. (2), with a transaction's rate $\lambda(s,a) = (1+a)^\beta \tilde{\lambda}_s$, where $\beta \in \mathbb{R}_+$. The β models the buyer's sensitivity to discounts. The larger the β, the more transactions will be attracted given the same discount. We set $\alpha = 0.001$ and $s_0 = 0$ by default. We also set $\tilde{\eta}_i = 1/(N_i(s,a)+1)$, $\epsilon = 0.1/(\tilde{N}_i(s)+1)$ and $Q^{(0)}(s,a) = 1$, where $N_i(s,a)$ and $\tilde{N}_i(s)$ denote the number of visiting (s,a) pair and state s up to i-th iteration.

Baselines and metrics. We compare our QLFP algorithm with: (1) Q-learning (Bradtke and Duff 1994), (2) speedy Q-learning (Azar et al. 2011), and (3) the case of not providing any discount. We do not compare with the Zap Q-learning (Devraj and Meyn 2017) because it needs to invert a square matrix of order $26 \cdot 10^6$ in each iteration, making it not practical to infer the optimal discount. We define the profit improvement of QLFP over the Q-learning as

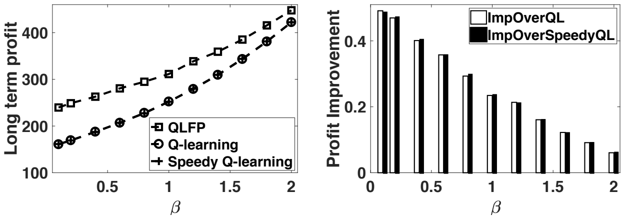
$$\text{ImpOverQL} \triangleq \frac{v^*(s|QLFP) - v^*(s|\text{Q-learning})}{v^*(s|\text{Q-learning})},$$

where $v^*(s|\text{Q-learning})$ denotes the long term profit under the Q-learning algorithm, i.e., Algorithm 1. Similarly, we define the improvement over speedy Q-learning and no discount as ImpOverSpeedyQL and ImpOverND respectively.

Impact of Demand

We study the impact of demand (i.e., parameter β). We consider the rating bias stated in Eq. (4). Figure 2 shows the long term profit and the profit improvement when β varies from 0.1 to 2. Figure 2(a) shows that the long term profit under

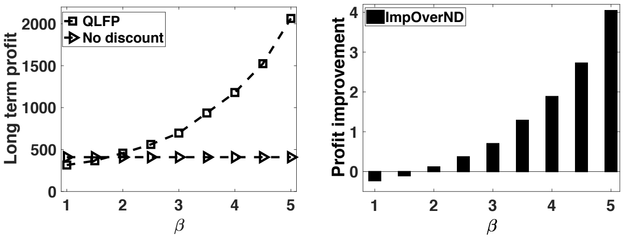
QLFP, Q-learning and speedy Q-learning increases as β increases (i.e., buyers become more sensitivity to discounts). Among these three algorithms, our QLFP algorithm has the largest long term profit. This implies that our QLFP converges faster than Q-learning and speedy Q-learning. From Figure 2(b), one can observe that the relative profit improvement is as high as 50%. The relative profit improvement decreases in β. Namely, the benefit of the forward projection decreases as buyers become more sensitive to discounts. This is because the forward projection preserves the monotonicity and its benefit is large when the $Q(s, a)$ is flat in s (i.e., when buyers are not very sensitive to discounts).



(a) Profit (b) Profit improvement

Figure 2: Impact of β on the profit and ImpOverQL and ImpOverSpeedyQL.

Figure 3 shows the long term profit and the profit improvement over no discount. Figure 3(a) shows that the long term profit is invariant of β when a seller does not provide any discount, while the long term profit under our QLFP algorithm increases significantly in β. Namely, using our QLFP algorithm, the sellers can earn more profit when buyers becomes more sensitive to discounts. Observe that when β is around 1 (i.e., buyers are not sensitive to discounts), our QLFP algorithm has a slightly smaller long term profit than the case of not providing any discount. This uncover a "*cost*" in inferring buyers' discount preferences from historical transaction data. When buyers are not sensitive to discounts, the cost of inference is larger than the benefit of providing discounts. This "*inference cost*" exists in general. Figure 3(b) shows that the profit improvement increases in β and the improvement can be as high as 4 times.



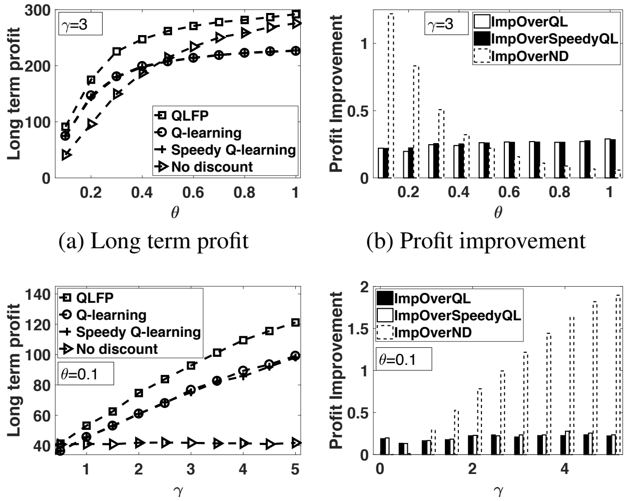
(a) Profit (b) Profit improvement

Figure 3: Impact of β on profit and ImpOverND.

Lessons learned: Our QLFP algorithm improves the profit over the Q-learning and Speedy Q-learning by as high as 50%, and over the case of not providing any price discount by as high as four times.

Impact of Rating Bias

Now we study the impact of rating bias (i.e., parameter θ and γ). We fix $\beta = 1$ and consider the rating bias stated in Eq. (5). Figure 4 and shows the long term profit and the profit improvement. Figure 4(a) and 4(d) show that the long term profit (of QLFP, Q-learning, speedy Q-learning and no discount) is non-decreasing in both θ and γ. This implies that the seller can earn more profit when buyers providing higher ratings. Furthermore, our QLFP has the largest long term profit among these four algorithms. From Figure 4(b) and Figure 4(e) one can observe that the profit improvement is as high as 30% over Q-learning and speedy Q-learning, and as high as two times over the case of no discount. This shows that our QLFP converges faster than Q-learning and speedy Q-learning and effective in inferring the optimal discount.



(a) Long term profit (b) Profit improvement

(c) Long term profit (d) Profit improvement

Figure 4: Impact of rating bias on profit and ImpOverQL and ImpOverSpeedyQL.

Related Work

Reputation systems (Resnick et al. 2000) are important in E-commerce systems. Several works investigated the economic efficiency of reputation systems in E-commerce applications. Dellarocas (Dellarocas 2001) studied the impact of rating leniency (from buyers) on sellers' advertising behavior. Khopkar *et al.* (Khopkar, Li, and Resnick 2005) studied the impact of negative feedback ratings on the efficiency of the eBay reputation system. Xie et al. (Xie and Lui 2015; 2017) formulated the "*ramp-up time*" to quantify the efficiency of reputation systems. Xie et al. (Xie, Ma, and Lui 2018) applied the stochastic bandit framework to select price discount online subjected to various trade-offs between the ramp up time and the short term profit. However, it is still unclear how the reputation effect compensates a seller's profit in the long run, i.e., the long term profit, and how to infer the optimal discount being aware of the long term profit. Out work applies a reinforcement learning approach to fill in this gap.

From an economic perspective, our work is related to (Landon and Smith 1998; Ba and Pavlou 2002; Jin and Kato 2006). Using a historical transaction dataset from the wine market, Landon *et al.* (Landon and Smith 1998) uncovered how the reputation of a wine influences its price. In online auction markets, Ba *et al.* (Ba and Pavlou 2002) found that a seller can have some price premiums if she has a high reputation, and Jin *et al.* (Jin and Kato 2006) studied how the reputation influences the pricing behavior of sellers in Internet auctions. We study a different problem, i.e., optimizing the reputation & discount trade-offs.

A variety of RL algorithms were designed for SMDP models (Bradtke and Duff 1994), such as the classical Q-learning, temporal difference learning, ATRDP and their variants. We refer the reader to (Bertsekas and Tsitsiklis 1996; Bradtke and Duff 1994; Sutton and Barto 1998) for a thorough treatment on RL. To infer the optimal discount, there are three notable Q-learning like algorithms, i.e., Q-learning (Bradtke and Duff 1994), Speedy Q-learning (Azar et al. 2011), and Zap Q-learning (Devraj and Meyn 2017). Our QLFP algorithm extends the classical Q-learning algorithm. We prove the convergence of our QLFP algorithm and show via experiments that our QLFP algorithm improves the profit by as high as 50% over both the classical Q-learning and speedy Q-learning algorithm. We do not compare with the Zap Q-learning algorithm because the it needs to invert a square matrix of order 26×10^6 in each iteration, making it not practical to infer the optimal discount.

Conclusion

This paper develops an online framework to optimize the reputation & discount trade-offs. We formulated a profit maximization problem via an SMDP to explore optimal trade-offs in selecting price discounts. We proved the monotonicity of the optimal profit and discount. Based on the monotonicity, we designed a QLFP algorithm, which infers the optimal discount from historical transaction data. We conducted experiments on a dataset from eBay to showed that our QLFP algorithm improves the profit by as high as 50% over the Q-learning and speedy Q-learning algorithm. Our QLFP algorithm also improves the profit by as high as four times over the case of not providing any discount.

Acknowledgments

The work of John C.S. Lui was supported in part by the GRF Funding 14200117. We would like to thank Richard T.B. Ma for his helpful discussions on the paper.

References

Azar, M. G.; Munos, R.; Ghavamzadeh, M.; and Kappen, H. 2011. Speedy q-learning. In *Advances in neural information processing systems*.

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

Bertsekas, D. P., and Tsitsiklis, J. N. 1996. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition.

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Bradtke, S. J., and Duff, M. O. 1994. Reinforcement learning methods for continuous-time markov decision problems. In *Proc. of NIPS*.

Chiang, A. C. 1984. Fundamental methods of mathematical economics.

Dellarocas, C. 2001. Analyzing the economic efficiency of ebay-like online reputation reporting mechanisms. In *Proc. of ACM EC*.

Devraj, A. M., and Meyn, S. 2017. Zap q-learning. In *Advances in Neural Information Processing Systems*, 2235–2244.

eBay. 1995. eBay Classifies Sellers into Twelve Stars. http://pages.ebay.com/help/feedback/scores-reputation.html.

Fortune500. 2015. http://fortune.com/fortune500/.

Jin, G. Z., and Kato, A. 2006. Price, quality, and reputation: Evidence from an online field experiment. *The RAND Journal of Economics* 37(4):983–1005.

Khopkar, T.; Li, X.; and Resnick, P. 2005. Self-selection, slipping, salvaging, slacking, and stoning: The impacts of negative feedback at ebay. In *Proc. of ACM EC*.

Landon, S., and Smith, C. E. 1998. Quality expectations, reputation, and price. *Southern Economic Journal* 628–647.

Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social influence bias: A randomized experiment. *Science* 341(6146):647–651.

Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Resnick, P.; Kuwabara, K.; Zeckhauser, R.; and Friedman, E. 2000. Reputation systems. *Commun. ACM* 43(12):45–48.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Xie, H., and Lui, J. C. S. 2015. Modeling ebay-like reputation systems: Analysis, characterization and insurance mechanism design. *Performance Evaluation* 91:132–149.

Xie, H., and Lui, J. C. S. 2017. Mining deficiencies of online reputation systems: Methodologies, experiments and implications. *IEEE Transactions on Services Computing*.

Xie, H.; Li, Y.; and Lui, J. C. 2018. *A Reinforcement Learning Approach to Optimize Discounts & Reputation Trade-offs in E-commerce Systems.* https://1drv.ms/b/s!AkqQNKuLPUbEdgLEiMLJQu8MfZM.

Xie, H.; Ma, R. T. B.; and Lui, J. C. S. 2018. Enhancing reputation via price discounts in e-commerce systems: A data-driven approach. *ACM Trans. Knowl. Discov. Data* 20(3):26:1–26:29.

Xu, H.; Liu, D.; Wang, H.; and Stavrou, A. 2015. E-commerce reputation manipulation: The emergence of reputation-escalation-as-a-service. In *Proc. of WWW*.