A Reinforcement Learning Approach to Optimize Discount and Reputation Tradeoffs in E-commerce Systems

HONG XIE, Chongqing University, China
YONGKUN LI, University of Science and Technology of China, China
JOHN C. S. LUI, The Chinese University of Hong, Hong Kong SAR

Feedback-based reputation systems are widely deployed in E-commerce systems. Evidence shows 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 as follows: (1) The demands from buyers depend on both the discount and reputation, and (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 to explore the optimal tradeoffs in selecting price discounts. We prove the monotonicity of the optimal profit and optimal discount. Based on the monotonicity, we design a Q-learning with forward projection (QLFP) algorithm, which infers the optimal discount from historical transaction data. We prove that the QLFP algorithm convergences to the optimal policy. We conduct trace-driven simulations using a dataset from eBay to evaluate the QLFP algorithm. Evaluation results show that QLFP improves the profit by as high as 50% over both Q-learning and Speedy Q-learning. The QLFP algorithm also improves both the reputation and profit by as high as two times over the scheme of not providing any price discount.

CCS Concepts: • **Information systems** → **Electronic commerce**; • **Computing methodologies** → *Machine learning algorithms*;

Additional Key Words and Phrases: Reputation systems, reinforcement learning, discount

ACM Reference format:
Hong Xie, Yongkun Li, and John C. S. Lui. 2020. A Reinforcement Learning Approach to Optimize Discount and Reputation Tradeoffs in E-commerce Systems. *ACM Trans. Internet Technol.* 20, 4, Article 37 (October 2020), 26 pages.
https://doi.org/10.1145/3400024

The work of Hong Xie was supported in part by National Nature Science Foundation of China (61902042), Chongqing High-Technology Innovation and Application Development Funds (cstc2019jscx-msxm0652, cstc2019jscx-fxyd0385). The work of John C. S. Lui was supported in part by GRF 14201819.

Authors' addresses: H. Xie, College of Computer Science, Chongqing University, No. 174 Shazhengjie, Shapingba, Chongqing, China; email: xiehong2018@cqu.edu.cn; Y. Li, School of Computer Science and Technology, University of Science and Technology of China, No. 96, JinZhai Road Baohe District, Hefei, Anhui, China; email: ykli@ustc.edu.cn; J. C. S. Lui, Department of Computer Science and Engineering, The Chinese University of Hong, Shatin N.T., Hong Kong, Hong Kong SAR; email: cslui@cse.cuhk.edu.hk.

1533-5399/2020/10-ART37 $15.00
https://doi.org/10.1145/3400024

1 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 [11] in terms of total revenue. This article considers 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, a feedback-based reputation system [23] is the most widely deployed, e.g., in eBay, Taobao, and so on. 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 1, while a neutral rating does not change the reputation score. To earn a four-star label (i.e., a reputable label), a seller must increase his or her reputation score to at least 500 [10].

Often, buyers are less willing to buy products from low-reputation sellers. The authors of Reference [29] found that new sellers need to spend at least 700 days (on average) to earn a reputable label. Hence, some sellers resort to "illegal means" to increase their reputation, i.e., authors in Reference [33] found that more than 11,000 sellers on 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 [33]. 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 his or her reputation score is no less than 500. A seller can attract 10 transactions per day if he or she is reputable; otherwise, he or 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, he or 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 he or she earns a reputable label, i.e., the price becomes 0.6, and he or she does not provide any discount after becoming reputable. Assume this discount increases the transaction volume to 2 per day. He or 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 that (1) price discounts can enhance sellers' reputation and (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*, and so on. This article studies the discount selection problem in such general settings, and we aim to answer the following: (1) How do we quantify the optimal tradeoffs in selecting price discounts? (2) How do we characterize the optimal profit and discount? (2) How do we perform online inference to determine the optimal discount from historical transaction data? Characterizing the optimal profit and discount would not only reveal fundamental understandings on the reputation and discount tradeoffs but also

offer insights to design algorithms to infer the optimal discount with a fast learning speed. Our contributions are as follows:

- 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, and so on. We formulate a profit maximization framework via an SMDP to quantify the optimal tradeoffs in determining the optimal price discount.
- We prove the monotonicity of the optimal profit and discount with respect to the reputation effect via *convex optimization* [4] and *comparative statics* [6].
- Based on the monotonicity of the optimal profit, we design a Q-learning with forward projection algorithm, which infers the optimal discount from historical transaction data. We prove that the QLFP algorithm converges to the optimal policy.
- We conduct trace-driven simulations using a dataset from eBay to evaluate the QLFP algorithm. Evaluation results show that QLFP improves the profit by as high as 50% over both Q-learning and Speedy Q-learning. The QLFP algorithm also improves both the reputation and profit by as high as two times over the scheme of not providing any price discount.

This article organizes as follows. Section 3 presents the system model and the problem formulation. Section 4 analyzes the optimal price discount. Section 5 presents the design of our QLFP algorithm. Section 6 presents the trace-driven simulation results using a dataset from eBay. Section 2 presents the related work. Section 7 presents the proofs to theorems and lemmas. We conclude in Section 8.

2 RELATED WORK

Reputation systems [23] are important in E-commerce systems. Two typical models have been proposed: (1) the feedback-based model [14, 26, 32, 35], which assesses reputation based on the feedbacks (or ratings) from users, and (2) the transitive trust model [12, 17, 24, 34], which assesses reputation based on the propagation of trust. The feedback-based model is widely deployed in E-commerce systems. However, reputation manipulations or inflations exist and a number of works have been proposed to address them. Xu et al. [33] uncovered a large volume of fake transactions in E-commerce systems. A number of algorithms to defend against fake feedbacks (or transactions) were proposed in References [9, 13, 34]. Miller et al. proposed a peer-prediction mechanism [20] to mitigate dishonest feedbacks. A nice survey can be found in Reference [13].

Several works investigated the economic efficiency of reputation systems in E-commerce applications. Dellarocas [7] studied the impact of rating leniency (from buyers) on sellers' advertising behavior. Khopkar et al. [18] studied the impact of negative feedback ratings on the efficiency of the eBay reputation system. Xie et al. [28–30] formulated the "ramp-up time" to quantify the efficiency of reputation systems. Xie et al. [31] applied the stochastic bandit framework to select price discount online subjected to various tradeoffs between the ramp up time and the short-term profit. Our work develops a unified framework to maximize the long-term profit and we introduce a new reinforcement learning framework, i.e., QLFP, to infer the optimal discount online.

From an economic perspective, our work is related to References [2, 16, 19]. Using a historical transaction dataset from the wine market, Landon et al. [19] uncovered how the reputation of a wine influences its price. In online auction markets, Ba et al. [2] found that a seller can have some price premiums if he or she has a high reputation, and Jin et al. [16] studied how the reputation influences the pricing behavior of sellers in Internet auctions. From a methodological perspective, our work is related to Markov Decision Processes [22] and reinforcement learning [5, 27]. We apply those techniques to study the reputation vs. discount tradeoffs in E-commerce applications.

A variety of RL algorithms were designed for SMDP models [5], such as classical Q-learning, temporal difference learning, ATRDP, and their variants. We refer the reader to References [3, 5,

27] for a thorough treatment on RL. This work considers using the RL approach to study the discount and reputation tradeoffs. We start from classical Q-learning rather than its sophisticated variants for the practical consideration that simplicity is important for real-world E-commerce applications. We show that the classical Q-learning can already improve profit significantly. Furthermore, inspired by our theoretical characterizations of the optimal profit, we design a QLFP algorithm to speed up the convergence of Q-learning. The projection idea developed in QLFP can also be applied to speed up the convergence of a variety of other RL algorithms. Through this we demonstrate how theoretical characterizations can be applied to improve RL algorithms. We believe that our work can inspire more effective reinforcement learning algorithms.

3 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 tradeoffs in selecting price discounts. Finally, we formulate an online discount selection problem, which infers the optimal discount from a seller's transaction data.

3.1 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, and so on. 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 who has a reputation score s; his or her reputation score becomes $s + m$ once he or 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. A seller's reputation score is set as $\hat{S}$ when it drops below $\hat{S}$, and it is set as S when it exceeds S.

3.2 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

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

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 $\Omega \subseteq \mathbb{R}$. We say X is larger than Y (written as $X \geq Y$), iff $\mathbb{P}[X > x] \geq \mathbb{P}[Y > x]$ holds for all $x \in \Omega$.

Definition 1 states that a random variable X is larger than Y if X is more likely to realize an outcome with a large value than Y. Note that $X \geq Y$ implies that $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

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

Assumption 1 captures (1) the herding behavior [21] that buyers give higher ratings to more reputable sellers and (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

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

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 $a \in \mathcal{A}$, $W(j, a) \geq W(s, a)$ holds for all $s > j$, where $s, j \in \mathcal{S}$. Given $s \in \mathcal{S}$, $W(s, b) \geq W(s, a)$ 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 and (2) the price effect that buyers are more willing to buy a product under a larger discount. Consider Equation (2); Assumption 2 means that $\lambda(s, a)$ increases in both s and a. One example of such $\lambda(s, a)$ is derived in Equation (3).

Assumption 3. *There exists two constants $\epsilon > 0, \delta > 0$ such that $F_W(\delta|s, a) \leq 1 - \epsilon$ 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. It eliminates the technical subtlety that an infinite number of transactions arriving within a finite time leads the profit to be infinite. In real-world E-commerce systems, Assumption 3 holds, because the number of transactions within a finite time is finite. Consider Equation (2); Assumption 3 means that $\lambda(s, a)$ is bounded, e.g., the $\lambda(s, a)$ expressed in Equation (3).

Modeling discount update. This article 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.

3.3 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 tradeoffs 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 ith 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, he or 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

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

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 tradeoff, 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

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

where s_i, a_i denote the reputation score and discount at the ith decision epoch and π denotes a policy [22], 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 his or her own historical transaction data and use his or her data to predict the optimal discount. Precisely, the $i$the transaction data item is associated with the following: (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, and (5) a rating denoted by m_i. For example, at the 0th 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 ith 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 ith 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 ith decision epoch, where $i = 0, 1, \ldots, \infty$,*

- *receive $\mathcal{H}_i$ and determine a discount a_i based on transaction history $\{\mathcal{H}_0, \mathcal{H}_1, \ldots, \mathcal{H}_i\}$,*

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

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

4 OPTIMAL OFFLINE DISCOUNTS

In this section, we study the offline discount selection problem. We first derive conditions that optimal discounts must satisfy. We draw implications from the model to characterize these optimality conditions. Finally, we characterize the optimal profit and apply techniques from *convex optimization* [4] and *comparative statics* [6] to characterize optimal discounts.

4.1 Optimality Conditions

Let $f_W(w|s,a)$ denote the probability density function of the inter-arrival time $W(s,a)$ of transactions

$$f_W(w|s,a) \triangleq \frac{\partial F_W(w|s,a)}{\partial w}, \qquad \forall w \in \mathbb{R}_+, s \in \mathcal{S}, a \in \mathcal{A}.$$

Given a seller is at state s and sets a discount a in the present decision epoch, let $\phi(s,a)$ denote the expected discount factor for the next decision epoch, formally

$$\phi(s,a) \triangleq \int_0^\infty e^{-\alpha w} f_W(w|s,a)dw, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

Note that $\phi(s,a) \in (0,1)$. We also call $\phi(s,a)$ the *per-epoch discount factor*. Economically, $\phi(s,a)$ can be interpreted as the inflation in one decision epoch, and the larger the $\phi(s,a)$, the smaller the inflation. Let $r(s,a)$ denote the *per-epoch discounted profit* (or reward), which can be expressed as

$$r(s,a) \triangleq \int_0^\infty k(s,a)e^{-\alpha w} f_W(w|s,a)dw = k(s,a)\phi(s,a), \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

Let $\pi^* = (d^*)^\infty$ denote an optimal SD policy. Note that if the seller has a reputation score s, then the optimal discount is $d^*(s)$. Let $v^*(s)$ denote the maximum long-term profit for a given initial reputation score $s_0 = s$. For example, $v^*(0)$ is the maximum long-term profit for a new seller. The maximum profit $v^*(s)$ is a unique solution of the following Bellman equations [22]:

$$v^*(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \phi(s,a) \sum_{j\in\mathcal{S}} p(j|s,a) v^*(j) \right\} = \max_{a\in\mathcal{A}} \left\{ \phi(s,a) V(s,a) \right\}, \qquad \forall s \in \mathcal{S},$$

where $V(s,a)$ is defined as

$$V(s,a) \triangleq k(s,a) + \sum_{j\in\mathcal{S}} p(j|s,a) v^*(j), \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

Note that the above expression is of a product form, because the profit $k(s,a)$ is delayed to be payed in the next decision epoch, i.e., when the next transaction arrives. Furthermore, the optimal policy (or optimal discount) satisfies

$$d^*(s) \in \arg\max_{a\in\mathcal{A}} \left\{ \phi(s,a) V(s,a) \right\}, \qquad \forall s \in \mathcal{S}.$$

The above optimality conditions enable us to derive a "scaling property" of the optimal discount.

Lemma 1. *Suppose we scale the system such that $\tilde{q} = \xi q$, $\tilde{c} = \xi c$, $\tilde{F}_W(w|s,a) = F_W(w|s,a)$, and $\tilde{F}_R(m|s,a) = F_R(m|s,a)$, where $\xi \in \mathbb{R}_+$. Then the optimal discount and long-term profit for the scaled system are $d^*(s)$ and $\xi v^*(s)$, respectively.*

Remark: Lemma 1 states that as we scale the selling price and product cost linearly at the same rate, the optimal discount remain the same and the optimal long-term profit is scaled linearly at the same rate. Therefore, we can simply normalize the selling price such that $q = 1$. This would help us simplify the discussion in experiments. Intuitively, Lemma 1 holds, because we only scale the price and cost linearly at the same rate, while keeping the other model parameters unchanged.

4.2 Implications from the Model

Based on the properties of the transaction's arrival process, we characterize the per-epoch discount factor $\phi(s,a)$ and the per-epoch discounted profit $r(s,a)$. These characterizations will enable us to further characterize the optimal discount.

Lemma 2. *We can derive $\phi(s,a)$ as*

$$\phi(s,a) = \alpha \int_0^\infty e^{-\alpha w} F_W(w|s,a)dw, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A},$$

where $\phi(s,a)$ is non-decreasing in s and a. If $F_W(w|s,a)$ is strictly concave in a, then $\phi(s,a)$ is strictly concave in a.

Remark: Lemma 2 states that the per-epoch discount factor should be larger (i.e., a smaller inflation in a decision epoch) if the seller has a higher reputation score or he or she sets a larger price discount. If $F_W(w|s,a)$ has a diminishing return in the price discount a, then the per-epoch discount factor has a diminishing return in a as well.

Lemma 3. *We can derive $r(s,a)$ as*

$$r(s,a) = \alpha\tilde{u}(a) \int_0^\infty e^{-\alpha w} F_W(w|s,a)dw, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

If $a \le u/q$, then $r(s,a)$ is non-decreasing in s; otherwise, $r(s,a)$ is non-increasing in s. If $a \in [u/q, 1]$, then $r(s,a)$ is non-increasing in a. If $F_W(w|s,a)$ is strictly concave in a, then $r(s,a)$ is strictly concave in $a \in (0, u/q)$.

Remark: Lemma 3 states that if the discount is not too large such that the per-unit profit $\tilde{u}(a)$ is positive, then the seller can earn profit per decision epoch as his or her reputation score increases; otherwise (i.e., per-unit profit $\tilde{u}(a)$ is negative), the seller will lose more profit per decision epoch. If $F_W(w|s,a)$ has a diminishing return in the price discount a, then the per-epoch discounted profit $r(s,a)$ has a diminishing return in a as well (provided $a \in (0, u/q)$). In general, when $a \in [0, u/q]$, $r(s,a)$ is not monotone in a, because the $\tilde{u}(a)$ is non-negative and non-increasing in a and $F_W(w|s,a)$ is non-decreasing in a.

The authors of Reference [30] found that the transactions in eBay follow a Poisson arrival process via analyzing a dataset from eBay. In the following corollary, we study the Poisson arrival of transactions.

Corollary 1. *Suppose $F_W(w|s,a)$ satisfies Equation (2). We have*

$$\phi(s,a) = \frac{\lambda(s,a)}{\lambda(s,a)+\alpha}, \quad r(s,a) = \frac{\lambda(s,a)\tilde{u}(a)}{\lambda(s,a)+\alpha}, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

If $\lambda(s,a)$ is strictly concave in a, then $\phi(s,a)$ is strictly concave in a and $r(s,a)$ is strictly concave in $a \in (0, u/q)$.

Remark: Corollary 1 states that given the Poisson arrival of transactions, if the arrival rate of transactions has a diminishing return in the price discount a, then both $\phi(s,a)$ and $r(s,a)$ have a diminishing return a. For example, Equation (3) satisfies the conditions in Corollary 1.

4.3 Optimal Profit and Discounts

It is mathematically intractable to derive the closed-form expression for the maximum long-term profit $v^*(s)$. 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. Suppose we increase the α to $\tilde{\alpha}$, i.e., $\tilde{\alpha} \geq \alpha$, while keeping all the other model parameters unchanged. Then $\tilde{v}^*(s) \leq v^*(s)$, where $\tilde{v}^*(\cdot)$ denotes the value function corresponding to $\tilde{\alpha}$.*

Remark: Theorem 1 states that the seller can earn more profit if his or her reputation score increases or the inflation decreases. In other words, sellers always have incentive to increase their reputation scores. Note that these monotone properties serve as an important building block for us to characterize the optimal discount. Intuitively, Theorem 1 is a consequence of the monotonicity of the rating and inter-arrival time of transactions with respect to the reputation s.

Definition 2. For each reputation score $s \in \mathcal{S}$, we define the associated action-dependent long-term profit as

$$Q(s,a) \triangleq \phi(s,a)V(s,a), \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

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 $d^*(s)$ maximizes $Q(s,a)$.

Lemma 4. *Suppose $a \in \mathcal{A}_s$, where $\mathcal{A}_s$ is defined as*

$$\mathcal{A}_s \triangleq \{a | Q(s,a) > 0, a \in \mathcal{A}\}, \qquad \forall s \in \mathcal{S}.$$

For all $j > \ell \geq s$, where $j, \ell, s \in \mathcal{S}$, $Q(j,a) \geq Q(\ell,a)$ holds.

Remark: Lemma 4 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 optimization problem to further study the optimal discount.

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

$$\begin{aligned} &\underset{a}{\text{maximize}} && \ln Q(s,a) = \ln \phi(s,a) + \ln V(s,a) \\ &\text{subject to} && a \in \mathcal{A}. \end{aligned}$$

Remark: 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 2. *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, implying a unique policy.*

Remark: Theorem 2 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, Equation (1) satisfies the condition on $F_R(m|s,a)$. Theorem 2 is a consequence of that the objective function $\ln Q(s,a)$ is strictly concave with respect to a.

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

Remark: Corollary 2 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, Equation (3) satisfies the condition on $\lambda(s,a)$.

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 the discount a as

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

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 3. *Suppose the conditions in Theorem 2 hold. If $h(s,a)$ is non-decreasing in α, then the unique optimal discount $d^*(s)$ is non-increasing in α. If $h(s,a)$ is non-decreasing in s, then the unique optimal discount $d^*(s)$ is non-increasing in s.*

Remark: Theorem 3 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. Furthermore, more reputable sellers set smaller discounts.

5 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. We prove the convergence of our QLFP algorithm.

ALGORITHM 1: Discount Selection Via Q-learning

Require: Discounting rate α, learning rate η_i, exploration probability ϵ_i, initialization $Q^{(0)}(s,a)$;

1: **for** $i = 1$ to ∞ **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: $Q^{(i)}(s_{i-1}, a_{i-1}) \leftarrow \hat{\phi}(s_{i-1}, a_{i-1}) \max_{a \in \mathcal{A}} Q^{(i-1)}(s_i, a) + \hat{r}(s_{i-1}, a_{i-1})$.
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 ϵ_i, $a_i \sim$ 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**

5.1 Q-learning for Online Discount Selection

We apply the Q-learning algorithm [5] 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 update the estimation of $Q(s,a)$ and then selects a discount based on the latest estimation of $Q(s,a)$. We formally outline this idea in Algorithm 1. To illustrate, suppose a seller is in the ith 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 ith 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. We next state that under mild assumptions on rating bias $F_R(m|s,a)$, Algorithm 1 converges to the optimal policy, i.e., it selects the optimal discount asymptotically.

Lemma 5. *Suppose $\hat{M}, M > 0$ and $F_R(m|s,a)$ is strictly increasing in m and assumption 3 holds. Suppose ϵ_i and η_i satisfy*

$$\lim_{i\to\infty} \tilde{\epsilon}_i(s,a) \to 0, \quad \sum_{i=0}^{\infty} \tilde{\epsilon}_i(s,a) = \infty, \quad \sum_{i=0}^{\infty} \tilde{\eta}_i(s,a) = \infty, \quad \sum_{i=0}^{\infty} \tilde{\eta}_i^2(s,a) < \infty,$$

where $\tilde{\epsilon}_i(s,a) \triangleq \epsilon_i, \tilde{\eta}_i(s,a) \triangleq \eta_i$ if $(s_i,a_i) = (s,a)$; otherwise, $\tilde{\epsilon}_i(s,a) \triangleq 0, \tilde{\eta}_i(s,a) \triangleq 0$. The estimation $Q^{(i)}(s,a)$ produced by Algorithm 1 converges to $Q(s,a), \forall s,a$, almost surely (i.e., with probability 1) as i goes to infinity.

Lemma 5 implies that when sellers' reputation goes up and down, Algorithm 1 can converge to an optimal policy, i.e., the optimal action will be selected. The condition on $F_R(m|s,a)$ is to guarantee that each state action pair will be visited infinitely often as i goes to infinite. Selecting η_i an ϵ_i is based on the stochastic approximation theory [25], which is a standard result.

5.2 Q-learning with Forward Projection

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 4 states that given a reputation score s and a discount a, if $Q(s,a) \geq 0$, then $Q(j+1,a) \geq Q(j,a)$ holds 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 4. Step 2 executes steps 2–7 of Algorithm 1 to obtain an estimation of $Q^{(i)}(s,a)$ based on $\{\mathcal{H}_0, \mathcal{H}_1, \ldots, \mathcal{H}_i\}$. Steps 3–7 makes the $Q^{(i)}(s,a)$ to satisfy Theorem 4 via *forward projection*, i.e., propagate the value of $Q^{(i)}(s_{i-1},a_{i-1})$ upward in terms of the reputation score. Our QLFP algorithm improves Algorithm 1, because this forward projection increases the accuracy of estimating $Q(s,a)$. We next prove the convergence of Algorithm 2.

ALGORITHM 2: QLFP Algorithm

Require: α, η_i, ϵ_i, $Q^{(0)}(s,a)$ (satisfies Theorem 4);
1: **for** $i = 0, 1$ to ∞ **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** $Q^{(i)}(j,a_{i-1}) < Q^{(i)}(j-1,a_{i-1})$ **then**
6: $Q^{(i)}(j,a_{i-1}) \leftarrow Q^{(i)}(j-1,a_{i-1})$.
7: **end if**
8: **end for**
9: **end if**
10: Execute step 8 of Algorithm 1.
11: **end for**

Theorem 4. *Consider the same assumptions and conditions in Lemma 5. The estimation $Q^{(i)}(s,a)$ produced by Algorithm 2 converges to $Q(s,a), \forall s,a$, almost surely as i goes to infinity.*

QLFP converges, because the forward projection makes the Q satisfies the monotone property, and thus it does not disturb the convergence of the Q function. Furthermore, QLFP converges

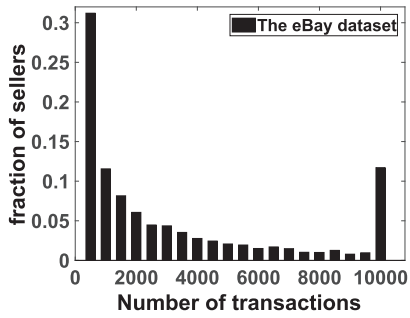


Fig. 1. The distribution of number of transactions.

Table 1. Reputation Score vs. the Number of Stars

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

faster than the Q-learning, because the forward projection utilizes the monotone property of the Q function to improve the accuracy $Q^{(i)}$ in estimating Q. The forward projection idea is generic and can be applied to speed up the convergence of other algorithms like variants of Q-learning, adaptive real-time dynamic programming [5], and so on. Similarly with the isotonic regression, the QLDP algorithm aims to learn a monotonic objective function. Unlike the isotonic regression, we consider the bandit feedback setting, i.e., the data arrive sequentially and the price discount influences subsequent data. Under this bandit setting, one needs to balance the exploration vs. exploitation tradeoff.

6 TRACE-DRIVEN EVALUATION

We evaluate the QLFP algorithm on a dataset from eBay. Evaluation results show that the QLFP improves the profit by as high as 50% over Q-learning and Speedy Q-learning and by as high as four times over the scheme of not providing any price discount.

6.1 Evaluation Settings

Datasets. We use a dataset from eBay [30], 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 timestamp, 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.

Model parameters. To assist buyers to assess sellers' reputation, eBay adopts a 12-star label system [10] summarized in Table 1. The authors of Reference [30] 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,

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

Table 2. Transaction's Rate across Number of Stars

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

Table 3. The Frequency of Each Rating

Rating	−1	0	1
Fraction	0.23%	0.34%	99.43%

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.,

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

In eBay, sellers with a reputation score or 10^6 or above have the same number of stars, i.e., 12 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 eBay adopts a three-level rating metric, i.e., {−1, 0, 1}. Thus, we set $\mathcal{M} = \{-1, 0, 1\}$. Table 3 summarizes the fraction of each rating level in our dataset. From Table 3, one can observe that 99.43% ratings are positive. Namely, the variation in the feedback rating is very small. This implies a very small bias in providing feedback ratings. Thus, we set the rating distribution as the empirical fraction of rating levels, i.e.,

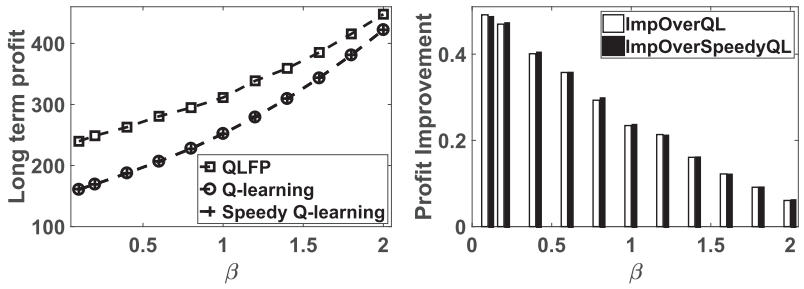
$$F_R(-1|s,a) = 0.0023, F_R(0|s,a) = 0.0057, F_R(1|s,a) = 1, \tag{4}$$

holds for all $s \in \mathcal{S}, a \in \mathcal{A}$. This treatment of the rating distribution is model free and does not introduce parameter tuning. It is suitable only when the rating bias is small, else one needs to sophisticated models.

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

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

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



(a) Profit (b) Profit improvement

Fig. 2. Impact of β on the profit and ImpOverQL and ImpOverSpeedyQL.

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.

By the linear-scaling property (i.e., Lemma 1), we normalize the baseline price to be $q = 1$ without any loss of generality. Furthermore, we set the cost and the discount set to be $c = 0.6, \mathcal{A} = \{0.02k | k = 0, 1, \ldots, 25\}$. Note that in this section, we focus on demonstrating how to apply our method to set discount for one product. Through this, we can simplify the presentation. If a seller has multiple products, then he or she may use our method to set discount for each product individually. With discounts, we still consider a Poisson arrival of transactions, i.e., $F_W(w|s, a)$ satisfies Equation (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 a discounting factor $\alpha = 0.001$ and initial state $s_0 = 0$ by default. We set a learning rate $\tilde{\eta}_i = 1/(N_i(s, a) + 1)$, an exploration probability $\epsilon = 0.1/(\tilde{N}_i(s) + 1)$ and an initialization $Q^{(0)}(s, a) = 1$, where $N_i(s, a)$ and $\tilde{N}_i(s)$ denote the number of visiting (s, a) pair and state s up to ith iteration.

We use the above parameters to simulate the model and run Algorithm 1 and 2 to predict discounts. Through this we obtain one sample of the long-term profit. We simulate the model for multiple rounds and use the sample average to estimate the long-term profit.

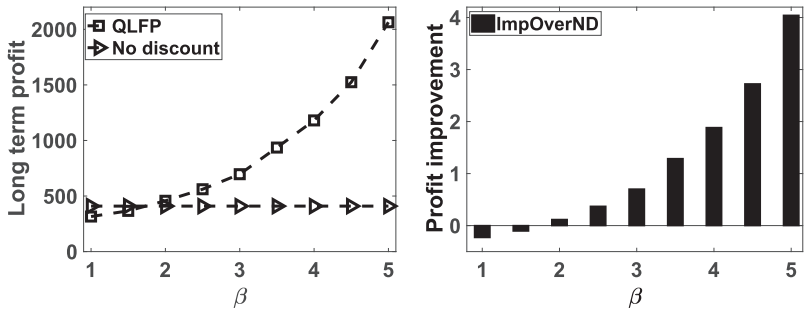
Baselines and metrics. We compare our QLFP algorithm with: (1) Q-learning [5], (2) Speedy Q-learning [1], and (3) the case of not providing any discount. We do not compare with the Zap Q-learning [8], because it needs to invert a square matrix of order $26 \cdot 10^6$ in each iteration, making it not 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.

6.2 Impact of Demand

We study the impact of demand (i.e., parameter β). We consider the rating bias stated in Equation (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. The reason is that the QLFP



(a) Profit (b) Profit improvement

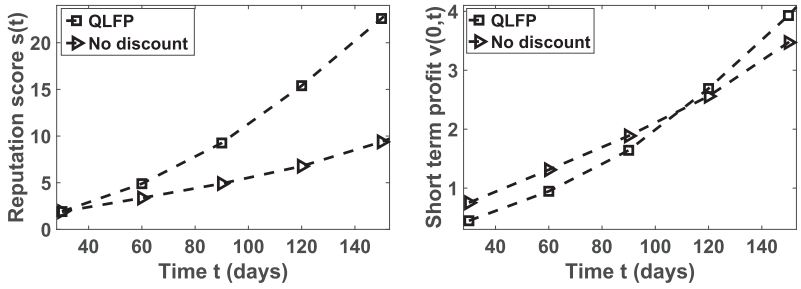
Fig. 3. Impact of β on profit and ImpOverND.

algorithm preserves the monotonicity of the Q function in each decision epoch. Preserving this monotonicity can improve the estimation of Q more accurate, and thus the QLFP algorithm can estimate the optimal discount with a higher accuracy. 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).

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 uncovers a "cost" in inferring buyers' discount preferences from historical transaction data. Note that to balance the exploration vs. exploitation tradeoff, QLFP needs to selects some sub-optimal discounts, resulting in small profit in such decision epochs. The benefit the QLFP is that it selects discount dynamically based on the state, which takes buyers behavior into consideration. When buyers are not sensitive to discounts (i.e., setting a discount only slightly improves the transaction volume), the cost of exploration (i.e., small profit in exploration decision epochs) is larger than the benefit of providing discounts. Figure 3(b) shows that the profit improvement increases in β and the improvement can be as high as 4 times.

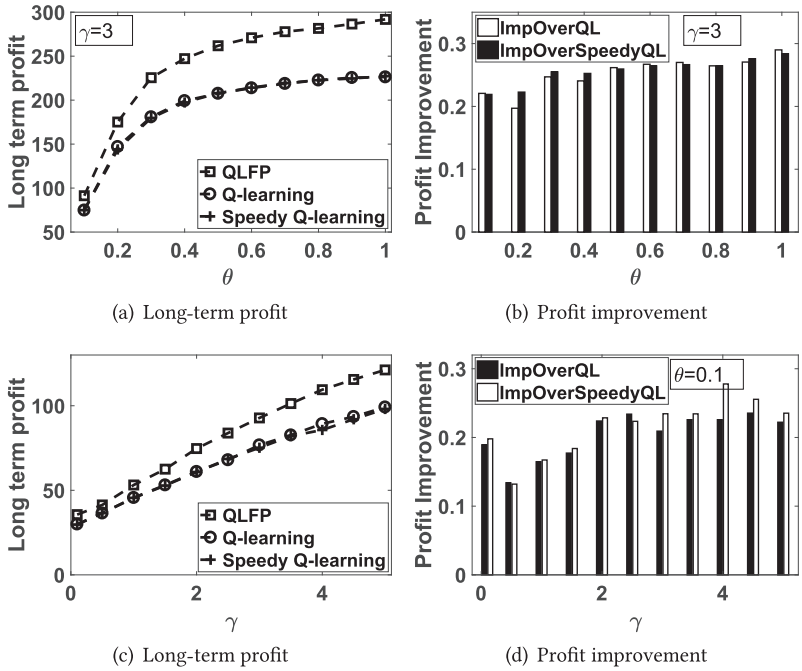
Now we study the reputation improvement and short-term profit loss of our QLFP algorithm. Let $s(t)$ and $v(s_0, t)$ denote the expected reputation score and discounted profit up to time $t \in \mathbb{R}_+$. Figure 4 shows $s(t)$ and $v(0, t)$ as the time t varies from 0 to 1,000 days, where we set $\beta = 3$. From Figure 4(a), one can observe that given the same amount of time our QLFP algorithm can earn a higher reputation score than the case of not providing any discount. In particular, when $t = 1,000$, the reputation score under our QLFP algorithm is around three times of that under no discounts. This reputation score improvement becomes larger as time t increases. Figure 4(b) shows that our QLFP algorithm loses a very small amount of short-term profit compared to the long-term profit. This shows that the seller will not be in risk of losing a large amount of short-term profit under the QLFP algorithm.

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. Furthermore, it improves the reputation by two times over the case of not providing any discount while loses a small amount of profit in a short term.



(a) Reputation score (b) Short-term profit

Fig. 4. The reputation and short-term profit.



(a) Long-term profit (b) Profit improvement

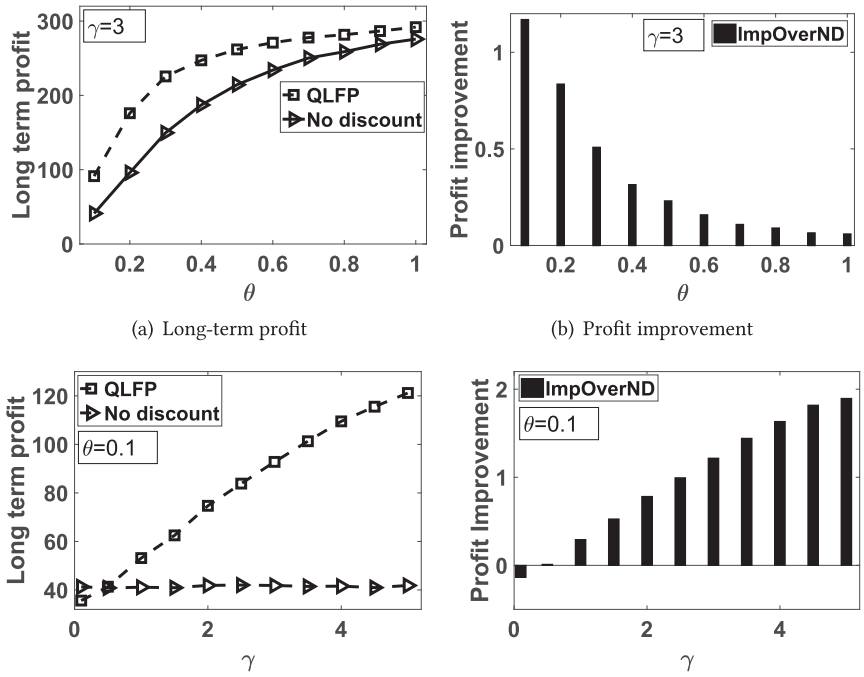
(c) Long-term profit (d) Profit improvement

Fig. 5. Impact of rating bias on profit and ImpOverQL and ImpOverSpeedyQL.

6.3 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 Equation (5). Figure 5 shows the long-term profit (of QLFP, Q-learning, and Speedy Q-learning) and the profit improvement. Figure 5(a) and and (c) shows that the long-term profit (of QLFP, Q-learning, and Speedy Q-learning) increases 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 three algorithms. From Figure 5(b) and (d), one can observe that the profit improvement is as high as 30%. This further justifies that our QLFP converges faster than Q-learning and Speedy Q-learning.

Figure 6 shows the long-term profit (for QLFP and no discount case) and the profit improvement. From Figure 6(a), one can observe that the long-term profit increases in θ. Namely, the



(a) Long-term profit (b) Profit improvement

(c) Long-term profit (d) Profit improvement

Fig. 6. Impact of rating bias on profit and ImpOverND.

seller can earn more profit when buyers provide higher ratings, i.e., becomes less bias. The curve corresponding to our QLFP algorithm lies above the case of not providing any discount. Namely, our QLFP algorithm improves the long-term profit over no discounts even under *large rating bias*. From Figure 6(b), one can observe that this improvement drops as θ increases. Figure 6(c) shows that the long-term profit under no discounts is invariant of γ and the long-term profit under our QLFP algorithm increases in γ. Namely, the seller can earn more profit (using our QLFP algorithm) when buyers' leniency is more sensitive to discount. Figure 6(d) shows that our QLFP algorithm can improve the long-term profit over no discounts by as high as 200% and this improvement increases as buyers' leniency becomes more sensitive to discount.

Lessons learned: Our QLFP algorithm improves the long-term profit over the no discount case, Q-learning, and Speedy Q-learning under different levels of rating biases.

7 PROOF FOR LEMMAS AND THEOREMS

Proof of Lemma 1. Let $\tilde{k}(s,a)$ denote the lump sum reward for the scaled system. Then, we have $\tilde{k}(s,a) = \tilde{q} - \tilde{c} - a\tilde{q} = \xi(q - c - aq) = \xi k(s,a)$. Let $\tilde{\phi}(s,a)$ denote expected discount rate. Then, we have $\tilde{\phi}(s,a) = \alpha \int_0^\infty e^{-\alpha w} \tilde{F}_W(w|s,a)dw = \alpha \int_0^\infty e^{-\alpha w} F_W(w|s,a)dw = \phi(s,a)$. Let $\tilde{r}(s,a)$ denote the per-epoch discounted profits for the scaled system. Then, we have that $\tilde{r}(s,a) = \tilde{k}(s,a)\tilde{\phi}(s,a) = \xi r(s,a)$. Let $\tilde{p}(s,a)$ denote the state transition probability for the scaled system. Then, we have

$$\tilde{p}(j|s,a) = \tilde{F}_R(j-s|s,a) - \tilde{F}_R(j-s-1|s,a) = F_R(j-s|s,a) - F_R(j-s-1|s,a) = p(j|s,a).$$

Then, we verify that $\tilde{v}^*(s) = \xi v^*(s)$ satisfies the optimality condition:

$$\tilde{r}(s,a) + \delta \sum_{j\in\mathcal{S}} \tilde{p}(j|s,a)\tilde{v}^*(j) = \xi r(s,a) + \delta \sum_{j\in\mathcal{S}} p(j|s,a)\xi v^*(s)$$
$$= \xi \left[r(s,a) + \delta \sum_{j\in\mathcal{S}} p(j|s,a)v^*(s) \right] = \tilde{v}^*(s).$$

Last, we verify that $\tilde{d}^*(s) = d^*(s)$ satisfies the optimality condition:

$$\arg\max_{a\in\mathcal{A}} \left\{ \tilde{r}(s,a) + \delta \sum_{j\in\mathcal{S}} \tilde{p}(j|s,a)\tilde{v}^*(j) \right\} = \arg\max_{a\in\mathcal{A}} \xi \left\{ r(s,a) + \delta \sum_{j\in\mathcal{S}} p(j|s,a)v^*(j) \right\} = d^*(s) = \tilde{d}^*(s).$$

This proof is then complete. □

Proof of Lemma 2. First we have that

$$\frac{\partial F_W(w|s,a)e^{-\alpha w}}{\partial w} = f_W(w|s,a)e^{-\alpha w} - \alpha F_W(w|s,a)e^{-\alpha w}.$$

Note that

$$\int_0^\infty \frac{\partial F_W(w|s,a)e^{-\alpha w}}{\partial w} dw = F_W(w|s,a)e^{-\alpha w}|_0^\infty = 0.$$

Thus, we have that

$$\phi(s,a) = \alpha \int_0^\infty e^{-\alpha w} F_W(w|s,a)dw.$$

Note that Assumption 2 implies that $F_W(w|s,a)$ is non-decreasing in s and a. Thus, $\phi(s,a)$ is non-decreasing in s and a. We can derive the second-order derivative of $\phi(s,a)$ with respect to a as

$$\frac{\partial^2\phi(s,a)}{\partial a^2} = \alpha \int_0^\infty e^{-\alpha w} \frac{\partial^2 F_W(w|s,a)}{\partial a} dw < 0.$$

We thus conclude the concavity of $\phi(s,a)$ with respect to a. □

Proof of Lemma 3. We can derive $r(s,a)$ as

$$r(s,a) = k(s,a)\phi(s,a) = \tilde{u}(a)\phi(s,a) = \alpha\tilde{u}(a) \int_0^\infty e^{-\alpha w} F_W(w|s,a)dw.$$

The remaining proof is similar to that of Lemma 2. □

Proof of Corollary 1. Note that $F_W(w|s,a) = 1 - e^{-\lambda(s,a)w}$. Then we can derive $\phi(s,a)$ as

$$\phi(s,a) = \alpha \int_0^\infty e^{-\alpha w} F_W(w|s,a)dw = \alpha \int_0^\infty e^{-\alpha w}(1 - e^{-\lambda(s,a)w})dw$$
$$= \alpha \int_0^\infty e^{-\alpha w}dw - \alpha \int_0^\infty e^{-\alpha w} e^{-\lambda(s,a)w}dw = 1 - \frac{\alpha}{\lambda(s,a)+\alpha} = \frac{\lambda(s,a)}{\lambda(s,a)+\alpha}.$$

With a similar derivation, we obtain $r(s,a)$. To conclude the concavity, we derive the second-order derivative as

$$\frac{\partial^2\phi(s,a)}{\partial a^2} = -\frac{2\alpha}{(\lambda(s,a)+\alpha)^3} \left(\frac{\partial\lambda(s,a)}{\partial a} \right)^2 + \frac{\alpha}{(\lambda(s,a)+\alpha)^2} \frac{\partial^2\lambda(s,a)}{\partial a^2} < 0.$$

Thus, we conclude the concavity of $\phi(s,a)$ with respect to a. To study the concavity of $r(s,a)$, we use the fact that $r(s,a) = \tilde{u}(a)\lambda(s,a)$. Note that the first and second derivatives of $\tilde{u}(a)$ are $\tilde{u}'(a) = -q$ and $\tilde{u}''(a) = 0$. Then, we derive the second-order derivative of $r(s,a)$ with respect to a as

$$\frac{\partial^2 r(s,a)}{\partial a^2} = \tilde{u}''(a)\lambda(s,a) + 2\tilde{u}'(a)\frac{\partial \lambda(s,a)}{\partial a} + \tilde{u}(a)\frac{\partial^2 \lambda(s,a)}{\partial a^2} = \tilde{u}(a)\frac{\partial^2 \lambda(s,a)}{\partial a^2} - 2q\frac{\partial \lambda(s,a)}{\partial a} < 0.$$

This proof is then complete. □

Proof of Theorem 1. We prove this theorem by induction. Given initial state $s_0 = s$, let $v_i^*(s)$ denote the maximum expected discounted profits up to the first ith decision epochs. Then, $v_0^*(s) = 0$, and $v_i^*(s)$ converges to the long-term profits $v^*(s)$ as i goes to infinity, i.e.,

$$v^*(s) = \lim_{i\to\infty} v_i^*(s).$$

To conclude this theorem, it suffices to show that for each $i = 0, 1, \ldots, \infty$, $v_i^*(s)$ is non-decreasing in s. We next show this by induction.

Consider the case $i = 0$. Note that $v_0^*(s) = 0$ for all s. Thus $v_0^*(s)$ is non-decreasing in s and $v_0^*(s) \geq 0$. Suppose that $v_i^*(s)$ is non-decreasing in s and $v_i^*(s) \geq 0, \forall s \in \mathcal{S}$. Based on it, we need to show that $v_{i+1}^*(s)$ is non-decreasing in s and $v_{i+1}^*(s) \geq 0, \forall s \in \mathcal{S}$. Observe that

$$v_{i+1}^*(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \phi(s,a) \sum_{j\in\mathcal{S}} p(j|s,a) v_i^*(j) \right\} = \max_{a\in\mathcal{A}} \left\{ \phi(s,a) V_i(s,a) \right\},$$

where $V_i(s,a)$ is defined as $V_i(s,a) \triangleq k(s,a) + \sum_{j\in\mathcal{S}} p(j|s,a) v_i^*(j)$. Let $\bar{F}_R(m|s,a) \triangleq 1 - F_R(m|s,a)$ denote the complementary CDF of R. Consider the case that $s - \hat{M} \geq -\hat{S}$ and $s + M \leq S$. We can derive $V_i(s,a)$ as

$$\begin{aligned} V_i(s,a) &= k(s,a) + \sum_{j=s-\hat{M}}^{s+M} p(j|s,a) v_i^*(j) = \tilde{u}(a) + \sum_{j=s-\hat{M}}^{s+M} [\bar{F}_R(j-s|s,a) - \bar{F}_R(j-s+1|s,a)] v_i^*(j) \\ &= \tilde{u}(a) + \sum_{j=s-\hat{M}+1}^{s+M} \bar{F}_R(j-s|s,a)[v_i^*(j+1) - v_i^*(j)] + v_i^*(s-\hat{M}). \end{aligned}$$

Assumption 1 implies that $F_R(m|s,a)$ is non-increasing in s. Given $\tilde{s} > s$, then it follows that

$$\bar{F}_R(j-\tilde{s}|\tilde{s},a) \geq F_R(j-s|\tilde{s},a) = 1 - F_R(j-s|\tilde{s},a) \geq 1 - F_R(j-s|s,a) = \bar{F}_R(j-s|s,a).$$

Note that $v_i^*(j+1) \geq v_i^*(j)$. Then it follows that the term $\sum_{j=s-\hat{M}+1}^{s+M} \bar{F}_R(j-s|s,a)[v_i^*(j+1) - v_i^*(j)]$ is non-decreasing in s for all s satisfies $s - \hat{M} \geq -\hat{S}$ and $s + M \leq S$. Note that $v_i^*(s-\hat{M})$ is non-decreasing in s for all s satisfies $s - \hat{M} \geq -\hat{S}$ and $s + M \leq S$. Thus $V_i(s,a)$ is non-decreasing in s for all s satisfies $s - \hat{M} \geq -\hat{S}$ and $s + M \leq S$. With a similar derivation as the above, we can extend this monotone property of $V_i(s,a)$ to the case $s - \hat{M} \leq -\hat{S}$ and the case $s + M \geq S$. Namely, $V_i(s,a)$ is non-decreasing in s for all $s \in \mathcal{S}$. Recall that $\phi(s,a)$ is non-decreasing in s for all $s \in \mathcal{S}$. To maximize $\phi(s,a)V_i(s,a)$, the discount must satisfy that $V_i(s,a) \geq 0$. This is because $V_i(s,0) \geq 0$ and $\lambda(s,a) > 0$. Let $d_i^*(s)$ denote the optimal discount to maximize $v_i^*(s)$. Then it follows that

$$\begin{aligned} v_{i+1}^*(s+1) &= \max_{a\in\mathcal{A}} \{\phi(s+1,a)V_i(s+1,a)\} \geq \phi(s+1, d_{i+1}^*(s)) V_i(s+1, d_{i+1}^*(s)) \\ &\geq \phi(s, d_{i+1}^*(s)) V_i(s, d_{i+1}^*(s)) = v_{i+1}^*(s). \end{aligned}$$

Thus, we conclude the monotonicity of $v^*(s)$ in s.

Consider another system with $\tilde{\alpha} \geq \alpha$. Let $\tilde{v}^*(s)$ denote the maximum profits corresponding to the system with $\tilde{\alpha}$. Our objective is to show $\tilde{v}^*(s) \leq v^*(s)$. Let $\tilde{v}_i^*(s)$ and $v_i^*(s)$ denote the maximum profits up to the ith decision epoch. It suffices to show that $\tilde{v}_i^*(s) \leq v_i^*(s)$ holds for all $i = 0, 1, \ldots, \infty$, because

$$\tilde{v}_i^*(s) \leq v_i^*(s) \Rightarrow \lim_{i\to\infty} \tilde{v}_i^*(s) \leq \lim_{i\to\infty} v_i^*(s) \Rightarrow \tilde{v}^*(s) \leq v^*(s).$$

We prove $\tilde{v}_i^*(s) \leq v_i^*(s)$ via induction. Consider the case that $i = 0$. Then, we have that $\tilde{v}_i^*(s) = 0$ and $v_i^*(s) = 0$ for all $s \in \mathcal{S}$. Namely, $\tilde{v}_i^*(s) \leq v_i^*(s)$ holds for all $s \in \mathcal{S}$. Suppose that $\tilde{v}_i^*(s) \leq v_i^*(s)$ holds for all $s \in \mathcal{S}$. It suffices to show $\tilde{v}_{i+1}^*(s) \leq v_{i+1}^*(s)$. Let $\tilde{d}_i^*(s)$ denote the optimal discount to maximize $\tilde{v}_i^*(s)$ and $\tilde{\phi}(s,a)$ denote the expected discount rate for the system with $\tilde{\alpha}$. Then, we have that $\tilde{v}_{i+1}^*(s) = \max_{a\in\mathcal{A}} \left\{\tilde{\phi}(s,a)\tilde{V}_i(s,a)\right\}$, where $\tilde{V}_i(s,a)$ is defined as $\tilde{V}_i(s,a) \triangleq k(s,a) + \sum_{j\in\mathcal{S}} p(j|s,a)\tilde{v}_i^*(j)$. Note that $\tilde{v}_i^*(s) \leq v_i^*(s)$. Thus, we have $\tilde{V}_i(s,a) \leq V_i(s,a)$. Observe that

$$\frac{\partial\phi(s,a)}{\partial a} = -\int_0^\infty we^{-\alpha w} f_W(w|s,a)dw < 0.$$

This implies that $\tilde{\phi}(s,a) < \phi(s,a)$. Note that $\tilde{V}_i(s,\tilde{d}_{i+1}^*(s)) > 0$ and $\phi(s,a) > 0$. Then it follows that

$$\begin{aligned}\tilde{v}_{i+1}^*(s) &= \tilde{\phi}(s,\tilde{d}_{i+1}^*(s))\tilde{V}_i(s,\tilde{d}_{i+1}^*(s)) \leq \phi(s,\tilde{d}_{i+1}^*(s))\tilde{V}_i(s,\tilde{d}_{i+1}^*(s))\\ &\leq \phi(s,\tilde{d}_{i+1}^*(s))V_i(s,\tilde{d}_{i+1}^*(s)) \leq \max_{a\in\mathcal{A}}\{\phi(s,a)V_i(s,a)\} = v_{i+1}^*(s).\end{aligned}$$

This proof is then complete. □

Proof of Lemma 4. Note that $Q(s,a) = \phi(s,a)V(s,a)$. Note that $\phi(s,a) > 0$. Thus, $V(s,a) \geq 0$ holds for all $a \in \mathcal{A}_s$. With a similar derivation as Theorem 1, we can obtain that $V(s,a)$ is non-decreasing in s. Then it follows that for each $a \in \mathcal{A}_s$, $V(j,a) \geq 0$ holds for all $j \geq s$, where $s, j \in \mathcal{S}$. Given $a \in \mathcal{A}_s$, for all $j > s$, where $s, j \in \mathcal{S}$, we have that

$$Q(j,a) = \phi(j,a)V(j,a) \geq \phi(s,a)V(j,a) \geq \phi(s,a)V(s,a) = Q(s,a).$$

This proof is then complete. □

Proof of Theorem 2. Note that the implicit constraint for Problem 3 is that $Q(s,a) > 0$. Namely, the feasible discounts for Problem 3 is $\mathcal{A}_s$. Let us first study the feasible discount set $\mathcal{A}_s$. Note that $\phi(s,a) > 0$. This implies that

$$Q(s,a) > 0 \Leftrightarrow \phi(s,a)V(s,a) > 0 \Leftrightarrow V(s,a) > 0.$$

Namely, $\mathcal{A}_s$ can be derived as $\mathcal{A}_s = \{a|V(s,a) > 0\}$. To further study $\mathcal{A}_s$, we next show that $V(s,a)$ is concave. Let us first consider the case that $s - \hat{M} \geq -\hat{S}$ and $s + M \leq S$. We can derive $V(s,a)$ as

$$\begin{aligned}V(s,a) &= k(s,a) + \sum_{j=s-\hat{M}}^{s+M} p(j|s,a)v^*(j) = \tilde{u}(a) + \sum_{j=s-\hat{M}}^{s+M} [F_R(j-s|s,a) - F_R(j-s-1|s,a)]v^*(j)\\ &= \tilde{u}(a) + \sum_{j=s-\hat{M}}^{s+M-1} F_R(j-s|s,a)[v^*(j) - v^*(j+1)] + v^*(s+M).\end{aligned}$$

Note that $\tilde{u}''(a) = 0$. Then we can derive the second-order derivative of $V(s,a)$ with respect to a as

$$\frac{\partial^2 V(s,a)}{\partial a^2} = \sum_{j=s-\hat{M}}^{s+M-1} \frac{\partial^2 F_R(j-s|s,a)}{\partial a^2}[v^*(j) - v^*(j+1)] \leq 0,$$

where the last inequality follows that $F_R(j-s|s,a)$ is convex in a, i.e., $\frac{\partial^2 F_R(j-s|s,a)}{\partial a^2} \geq 0$ and $v^*(s)$ is non-decreasing in s, i.e., $v^*(j) \leq v^*(j+1)$. Thus, $V(s,a)$ is concave with respect to a. For the case of $s - \hat{M} \leq -\hat{S}$ and the case of $s + M \geq S$, with a similar derivation, we can also have $V(s,a)$ is concave with respect to a. This implies that the set $\mathcal{A}_s = \{a|V(s,a) > 0\}$ is a convex set. Note that $V(s,0) > 0$. Then we can conclude that if $\{a|V(s,a) = 0, a \in [0,1]\} = \emptyset, \mathcal{A}_s = [0,1]$; otherwise, $\mathcal{A}_s = [0, \min\{a|V(s,a) = 0, a \in [0,1]\})$.

To conclude this theorem, the remaining thing is to show that the objective function $\ln Q(s,a)$ is strictly concave in $\mathcal{A}_s$. Let use derive the first-order derivative of $\ln Q(s,a)$ with respect to a as

$$\frac{\partial \ln Q(s,a)}{\partial a} = \frac{1}{\phi(s,a)} \frac{\partial \phi(s,a)}{\partial a} + \frac{1}{V(s,a)} \frac{\partial V(s,a)}{\partial a}.$$

Then the second-order derivative can be derived as

$$\frac{\partial^2 \ln Q(s,a)}{\partial a^2} = \frac{1}{\phi(s,a)} \frac{\partial^2 \phi(s,a)}{\partial a^2} - \frac{1}{\phi^2(s,a)} \left(\frac{\partial \phi(s,a)}{\partial a}\right)^2 + \frac{1}{V(s,a)} \frac{\partial^2 V(s,a)}{\partial a^2} - \frac{1}{V^2(s,a)} \left(\frac{\partial V(s,a)}{\partial a}\right)^2.$$

Note that $F_W(w|s,a)$ is strictly concave in a. Applying Theorem 2, we obtain that $\phi(s,a)$ is strictly concave in a, i.e., $\frac{\partial^2 \phi(s,a)}{\partial a^2} < 0$. We also have shown that $V(s,a)$ is concave with respect to a, i.e., $\frac{\partial^2 V(s,a)}{\partial a^2} \leq 0$. Observe that $\phi(s,a) > 0$ and $V(s,a) > 0$ hold for all $a \in \mathcal{A}_s$. We then conclude that $\frac{\partial^2 \ln Q(s,a)}{\partial a^2} < 0$, i.e., the objective function $\ln Q(s,a)$ is strictly concave in a. □

Proof of Corollary 2. Note that $F_W(w|s,a) = 1 - e^{-\lambda(s,a)w}$. Applying Theorem 2, it suffices to show that $F_W(w|s,a)$ is strictly concave with respect to a. Let us derive the second-order derivative with respect to a as

$$\frac{\partial^2 F_W(w|s,a)}{\partial a^2} = we^{-\lambda(s,a)w} \frac{\partial^2 \lambda(s,a)}{\partial a^2} - w^2 e^{-\lambda(s,a)w} \left[\frac{\partial \lambda(s,a)}{\partial a}\right]^2 < 0.$$

This proof is then complete. □

Proof of Theorem 3. Recall that in the Proof of Theorem 2, we showed that the feasible domain of Problem 3 is $\mathcal{A}_s$ and if $\{a|V(s,a) = 0, a \in [0,1]\} = \emptyset$, $\mathcal{A}_s = [0,1]$; otherwise, $\mathcal{A}_s = [0, \min\{a|V(s,a) = 0, a \in [0,1]\})$.

Let us first consider the case $\{a|V(s,a) = 0, a \in [0,1]\} = \emptyset$, i.e., $\mathcal{A}_s = [0,1]$. In this case, the optimal discount $d^*(s)$ falls into one of the three cases: (1) $d^*(s) = 0$, (2) $d^*(s) = 1$, (3) $d^*(s) \in (0,1)$ and $h(s, d^*(s)) = 0$. Note the the third case is actually the first-order optimality condition, because $h(s,a) = -\frac{\partial \ln Q(s,a)}{\partial a}$. Note that conditions in Theorem 2 hold. This means that $h(a)$ is strictly decreasing in $[0,1]$. Consider the third case $d^*(s) \in (0,1)$ and $h(s, d^*(s)) = 0$. Note that $\{a|V(s,a) = 0, a \in [0,1]\} = \emptyset$ implies that $\{a|V(s+1,a) = 0, a \in [0,1]\} = \emptyset$. Thus $\mathcal{A}_{s+1} = [0,1]$. Note that $h(s,a)$ is non-decreasing in s. Then we have that $h(s+1, d^*(s)) \geq h(s, d^*(s)) = 0$. Consider the case $h(s+1, d^*(s)) = 0$; we conclude that $d^*(s+1) = d^*(s)$ due to the concavity of $\ln Q(s,a)$. Consider the case $h(s+1, d^*(s)) > 0$. If the optimal discount $d^*(s+1)$ satisfies $h(s+1, d^*(s+1)) = 0$, then $d^*(s+1) < d^*(s)$ due to that $h(s+1,a)$ is non-decreasing in a. If $d^*(s+1)$ does not satisfy $h(s+1, d^*(s+1)) = 0$, then $h(s+1,a) > 0$ holds for all $a \in [0,1]$. Namely, $\frac{\partial \ln Q(s,a)}{\partial a} < 0$ holds for all $a \in [0,1]$. This implies that $d^*(s+1) = 0$, i.e., $d^*(s+1) \leq d^*(s)$. Thus, we conclude that $d^*(s+1) < d^*(s)$. Now consider the case $d^*(s) = 0$. This means that $h(s,a) > 0$ holds for all $a \in (0,1]$. Consider $a \in (0,1]$; we have that $h(s+1,a) \geq h(s,a) > 0$. This implies that $d^*(s+1) = 0$. Last, we consider the case $d^*(s) = 1$. This case is trivial, because $d^*(s) \leq 1$.

Now we consider the case $\mathcal{A}_s = [0, \min\{a|V(s,a) = 0, a \in [0,1]\})$. In this case, the optimal discount $d^*(s)$ falls into one of the following two cases: (1) $d^*(s) = 0$; (2) $d^*(s) \in [0, \min\{a|V(s,a) =$

$0, a \in [0,1]\}$) and $h(s, d^*(s)) = 0$. The proof is similar to the above case, i.e., $\{a|V(s,a) = 0, a \in [0,1]\} = \emptyset$. Finally, the above proof can be easily extended to prove that the optimal discount is non-increasing in α. □

Proof of Lemma 5. It suffices to prove that $Q^{(i)}(s,a)$ converges to $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}$. It follows from $\tilde{\epsilon}_i(s,a) > 0, \sum_{i=0}^{\infty} \tilde{\epsilon}_i(s,a) = \infty$ that each state action pair (s,a) will be visited infinity often. Precisely, for each (s,a), we can find $\tilde{\eta}_i(s,a) > 0$ such that $\sum_{i=0}^{\infty} \tilde{\eta}_i(s,a) = \infty$ and $\sum_{i=0}^{\infty} \tilde{\eta}_i^2(s,a) < \infty$. Note that $\hat{S}$ and S are finite. With a similar method as in Reference [3] this lemma can be concluded by showing that $\bar{Q}_i(s,a), \forall s, a, i$, is finite almost surely. Let a_{max} denote the maximum discount in $\mathcal{A}$. Consider the scenario that the arrival time of a reward follows the distribution $F(w|S, a_{max})$, and each time a reward $u+q$ is generated. Let w_i denote the waiting time of the ith reward and $\bar{Q}_0 = 0$. Furthermore, a reward is $\bar{Q}_{i+1} = (1-\tau_i)\bar{Q}_i + \tau_i(\bar{Q}_i e^{-\alpha w_{i+1}} + (u+q)e^{-\alpha w_{i+1}})$. Assumption 3 implies that $\mathbb{E}[e^{-\alpha w_{i+1}}] < 1$. Then, following some standard stochastic approximation argument, we have that $\bar{Q}_{i+1}$ converges almost surely to a finite value if $\tau_i > 0, \sum_{i=0}^{\infty} \tau_i = \infty$ and $\sum_{i=0}^{\infty} \tau_i^2 < \infty$ [25]. This means that $\bar{Q}_{i+1}$ is finite almost surely. □

Proof of Theorem 4. Note that from the proof of theorem 5, we know that $Q^{(i)}(s,a), \forall i, s, a$ are bounded almost surely. We extend the method developed in Reference [3] to prove this the convergence of $Q^{(i)}(s,a), \forall i, s, a$.

Note that $Q^{(i)}(s,a), \forall i, s, a$ are bounded almost surely, then there exists a D_0, such that $|Q^{(i)}(s,a) - Q(s,a)| < D_0$ holds for all i, s, a. We next show via induction that there exists $\tilde{i}$ and a constant $c < 1$ such that $|Q^{(i)}(s,a) - Q(s,a)| < cD_0$ holds for all $i \geq \tilde{i}, s, a$. Repeating this induction, we can make the diameter of $|Q^{(i)}(s,a) - Q(s,a)|$ shrinks to zero. Without loss of generality, we focus on one action a, and consider all the state-action pair $(-\hat{S}, a), \ldots, (S, a)$. Note that our forward projection algorithm, i.e., Algorithm 2, never updates the value of $Q^{(i)}(-\hat{S}, a)$, because the state $\hat{S}$ is the smallest possible state. Thus, with a similar method as that in Reference [3], we have that for any $c_1 > 0$, there exists $\hat{i}_1$ such that $|Q^{(i)}(-\hat{S}, a) - Q(-\hat{S}, a)| < (\phi(S, a_{max}) + c_1)D_0$ holds for all $i > \hat{i}_1$. Here c_1 is carefully selected such that $\phi(S, a_{max}) + c_1 < 1$.

Now we show that for any $c_2 > c_1$, there exists $\hat{i}_2$ such that $|Q^{(i)}(1-\hat{S}, a) - Q(1-\hat{S}, a)| < (\phi(S, a_{max}) + c_2)D_0$ holds for all $i \geq \hat{i}_2$. If the value of $Q^{(i)}(1-\hat{S}, a)$ is adjusted by forward projection finite number of times, then with a similar method as that in Reference [3] we have the desired result. Now let us consider $Q^{(i)}(1-\hat{S}, a)$ is adjusted by an infinite number of forward projections. Now we consider $i \geq \hat{i}_1$. For the ease of presentation, in the following proof, each subscript i represents $i + \hat{i}_1$. Let

$$w_i(1-\hat{S}, a) = \Big\{\hat{r}(1-\hat{S}, a) + \hat{\phi}(1-\hat{S}, a) \max_{a' \in \mathcal{A}} Q^{(i)}(s_{i+1}, a) - \mathbb{E}[\hat{r}(1-\hat{S}, a) + \hat{\phi}(1-\hat{S}, a) \max_{a' \in \mathcal{A}} Q^{(i)}(s_{i+1}, a)]\Big\} \times \mathbf{I}_{\{(s_i, a_i) = (1-\hat{S}, a)\}}.$$

Let us define the error sequence as

$$W_0(1-\hat{S}, a) = 0, W_{i+1}(1-\hat{S}, a) = [(1 - \tilde{\eta}_i(1-\hat{S}, a))W_i + \tilde{\eta}_i(1-\hat{S}, a)w_i(1-\hat{S}, a)]^+.$$

In the following proof, we will use the basic property that $W_i(\hat{S}+1, a)$ converges to 0 almost surely [15]. Let us define another sequence

$$Y_0 = D_0, Y_{i+1} = (1 - \tilde{\eta}_i(1-\hat{S}, a))Y_i + \tilde{\eta}_i(1-\hat{S}, a)(\phi(S, a_{max}) + c_2)D_0.$$

We only consider appropriate ϵ_2 such that $\phi(S, a_{max}) + c_2 < 1$. Then via induction, one can easily show that $Y_i \geq (\phi(S, a_{max}) + c_2)D_0$. Then we claim $Q^{(i+1)}(1-\hat{S}, a) < Q(1-\hat{S}, a) + Y_{i+1} + W_{i+1}(1-\hat{S}, a)$. We prove this claim via induction. Note that $Q^0(1-\hat{S}, a) \leq D_0 + Q(1-\hat{S}, a) =$

$Q(1-\hat{S},a)+Y_0+W_0(1-\hat{S},a)$. Suppose our claim hold for all $0,1,\ldots,i$. We next show it holds for $i+1$. Suppose $Q^{(i+1)}(1-\hat{S},a)$ is not updated, then $\tilde{\eta}_i(1-\hat{S},a)=0$. This implies that $Y_{i+1}=Y_i$ and $W_{i+1}=W_i$. Namely, our claim holds. Suppose we update $Q^{(i+1)}(1-\hat{S},a)$ via projection, then we have

$$\begin{aligned}Q^{(i+1)}(1-\hat{S},a) &\le Q(-\hat{S},a)+(\phi(S,a_{max})+c_1)D_0 \le Q(1-\hat{S},a)+(\phi(S,a_{max})+c_2)D_0\\ &\le Q(1-\hat{S},a)+Y_{i+1}+W_{i+1}(1-\hat{S},a).\end{aligned}$$

Now, suppose we update $\hat{Q}^{(i+1)}(1-\hat{S},a)$ is not via projection,

$$\begin{aligned}Q^{(i+1)}(1-\hat{S},a) &\le Q(1-\hat{S},a)+\tilde{\eta}_i(\hat{S}+1,a)\phi(S,a_{max})D_0+(1-\tilde{\eta}_i(1-\hat{S},a))(Y_i+W_i(1-\hat{S},a))\\ &\quad+\tilde{\eta}_i(\hat{S}+1,a)w_i(1-\hat{S},a)\\ &\le Q(1-\hat{S},a)+\tilde{\eta}_i(\hat{S}+1,a)(\phi(S,a_{max})+c_2)D_0+(1-\tilde{\eta}_i(1-\hat{S},a))Y_i\\ &\quad+(1-\tilde{\eta}_i(1-\hat{S},a))W_i(1-\hat{S},a)+\tilde{\eta}_i(\hat{S}+1,a)w_i(1-\hat{S},a)\\ &= Q(1-\hat{S},a)+Y_{i+1}+W_{i+1}(1-\hat{S},a).\end{aligned}$$

Note that $W_i(1-\hat{S},a)$ converges to 0 almost surely [15]. Also observe that Y_i converges to $(\phi(S,a_{max})+c_2)D_0$. Thus, we have

$$\limsup_{i\to\infty} Q^{(i)}(1-\hat{S},a) \le Q(1-\hat{S},a)+(\phi(S,a_{max})+c_2)D_0.$$

Note that the forward projection increases the value of $Q^{(i)}(\hat{S}+1,a)$ as compared to the case of without projection. For the case of without forward projection, implies a finite number of projection, thus a lower bound is

$$\liminf_{i\to\infty} Q^{(i)}(1-\hat{S},a) \ge Q(1-\hat{S},a)-\phi(S,a_{max})D_0.$$

Repeating this argument sequentially for $2-\hat{S},\ldots,S$, we have that there exists $\tilde{i}$ and $c<1$ such that $|Q^{(i)}(s,a)-\hat{Q}^*(s,a)|<cD_0$ for all $i>\tilde{i},s$. Finally, repeating the above whole argument for other actions, we complete this proof. □

8 CONCLUSION

This article develops an online framework to optimize the reputation & discount tradeoffs. We formulated a profit maximization problem via an SMDP to explore optimal tradeoffs 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 proved that the QLFP algorithm converges to the optimal policy. We conduct trace-driven simulations to evaluate the QLFP algorithm using a dataset from eBay. Evaluation results show that QLFP improves the profit by as high as 50% over both Q-learning and Speedy Q-learning. The QLFP algorithm also improves both the reputation and profit by as high as two times over the scheme of not providing any price discount.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments to improve this article.

REFERENCES

[1] Mohammad Gheshlaghi Azar, Remi Munos, Mohammad Ghavamzadeh, and Hilbert Kappen. 2011. Speedy Q-learning. In *Advances in Neural Information Processing Systems.*

[2] Sulin Ba and Paul A. Pavlou. 2002. Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quart.* 26, 3 (2002), 243–268.
[3] Dimitri P. Bertsekas and John N. Tsitsiklis. 1996. *Neuro-Dynamic Programming* (1st ed.). Athena Scientific.
[4] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.
[5] Steven J. Bradtke and Michael O. Duff. 1994. Reinforcement learning methods for continuous-time Markov decision problems. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS'94)*.
[6] Alpha C. Chiang. 1984. *Fundamental Methods of Mathematical Economics*. McGraw-Hill/Irwin, Boston, Mass.
[7] Chrysanthos Dellarocas. 2001. Analyzing the economic efficiency of eBay-like online reputation reporting mechanisms. In *Proceedings of the ACM Conference on Economics and Computation (EC'01)*.
[8] Adithya M. Devraj and Sean Meyn. 2017. Zap Q-learning. In *Advances in Neural Information Processing Systems*. 2235–2244.
[9] Prashant Dewan and Partha Dasgupta. 2010. P2P reputation management using distributed identities and decentralized recommendation chains. *IEEE Trans. Knowl. Data Eng.* 22, 7 (2010), 1000–1013.
[10] eBay. 1995. eBay Classifies Sellers into Twelve Stars. Retrieved from http://pages.ebay.com/help/feedback/scores-reputation.html.
[11] Fortune500. 2015. Retrieved from http://fortune.com/fortune500/.
[12] Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2004. Propagation of trust and distrust. In *Proceedings of the Annual Conference on the World Wide Web (WWW'04)*. 403–412.
[13] Kevin Hoffman, David Zage, and Cristina Nita-Rotaru. 2009. A survey of attack and defense techniques for reputation systems. *ACM Comput. Surv.* 42, 1, Article 1 (December 2009), 31 pages.
[14] Daniel Houser and John Wooders. 2006. Reputation in auctions: Theory, and evidence from eBay. *J. Econ. Manage. Strategy* 15, 2 (2006).
[15] Daniel R. Jiang and Warren B. Powell. 2015. An approximate dynamic programming algorithm for monotone value functions. *Operat. Res.* 63, 6 (2015), 1489–1511.
[16] Ginger Zhe Jin and Andrew Kato. 2006. Price, quality, and reputation: Evidence from an online field experiment. *AND J. Econ.* 37, 4 (2006), 983–1005.
[17] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. 2003. The eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the Annual Conference on the World Wide Web (WWW'03)*.
[18] Tapan Khopkar, Xin Li, and Paul Resnick. 2005. Self-selection, slipping, salvaging, slacking, and stoning: The impacts of negative feedback at eBay. In *Proceedings of the ACM Conference on Economics and Computation (EC'05)*.
[19] Stuart Landon and Constance E. Smith. 1998. Quality expectations, reputation, and price. *South. Econ. J.* 64, 3 (1998), 628–647.
[20] Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Manage. Sci.* 51, 9 (September 2005), 1359–1373.
[21] Lev Muchnik, Sinan Aral, and Sean J. Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.
[22] Martin L. Puterman. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.
[23] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (December 2000), 45–48.
[24] Paul Resnick and Rahul Sami. 2009. Sybilproof transitive trust protocols. In *Proceedings of the ACM Conference on Economics and Computation (EC'09)*.
[25] Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *Ann. Math. Stat.* 22, 3 (1951), 400–407.
[26] Aameek Singh and Ling Liu. 2003. TrustMe: Anonymous management of trust relationships in decentralized P2P systems. In *Proceedings of the Annual Peer-to-Peer Conference (P2P'03)*.
[27] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. Vol. 1. MIT press Cambridge.
[28] Hong Xie and John C. S. Lui. 2015. A data driven approach to uncover deficiencies in online reputation systems. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'15)*.
[29] Hong Xie and John C. S. Lui. 2015. Modeling eBay-like reputation systems: Analysis, characterization and insurance mechanism design. *Perf. Eval.* 91 (2015), 132–149.
[30] Hong Xie and John C. S. Lui. 2017. Mining deficiencies of online reputation systems: Methodologies, experiments and implications. *IEEE Trans. Serv. Comput.* 13, 5 (2017), 887–900. DOI: https://doi.org/10.1109/TSC.2017.2730206
[31] Hong Xie, Richard T. B. Ma, and John C. S. Lui. 2018. Enhancing reputation via price discounts in E-commerce systems: A data-driven approach. *ACM Trans. Knowl. Discov. Data* 20, 3, Article 26 (Jan. 2018), 29 pages. DOI: https://doi.org/10.1145/3154417
[32] Li Xiong and Ling Liu. 2004. Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Trans. Knowl. Data Eng.* 16, 7 (2004), 843–857.

[33] Haitao Xu, Daiping Liu, Haining Wang, and Angelos Stavrou. 2015. E-commerce reputation manipulation: The emergence of reputation-escalation-as-a-service. In *Proceedings of the Annual Conference on the World Wide Web (WWW'15)*.

[34] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. 2006. SybilGuard: Defending against sybil attacks via social networks. In *Proceedings of the ACM Special Interest Group on Data Communication Conference (SIGCOMM'06)*.

[35] Xiuzhen Zhang, Lishan Cui, and Yan Wang. 2014. Commtrust: Computing multi-dimensional trust by mining e-commerce feedback comments. *IEEE Trans. Knowl. Data Eng.* 26, 7 (2014), 1631–1643.

Received November 2019; revised April 2020; accepted May 2020