Competing Bandits:
The Perils of Exploration under Competition

Anonymous Author(s)

ABSTRACT

We empirically study the interplay between *exploration* and *competition*. Systems that learn from interactions with users often engage in *exploration*: making potentially suboptimal decisions in order to acquire new information for future decisions. However, exploration may hurt system's reputation in the near term, with adverse competitive effects. In particular, a system may enter a "death spiral" when decreasing market share leaves the system with less users to learn from, which degrades system's performance relative to competition and further decreases the market share.

We ask whether better exploration algorithms are incentivized under competition. We run extensive numerical experiments in a stylized duopoly model in which two firms deploy multi-armed bandit algorithms and compete for myopic users. We find that duopoly and monopoly tend to favor a primitive "greedy algorithm" that does not explore and leads to low consumer welfare, whereas a temporary monopoly (a duopoly with an early entrant) may incentivize better bandit algorithms and lead to higher consumer welfare. Our findings shed light on the first-mover advantage in the digital economy by exploring the role that data can play as a barrier to entry in online markets.

CCS CONCEPTS

• **Computing methodologies → Online learning settings**; *Sequential decision making*; • **Applied computing → Economics**;

KEYWORDS

Multi-armed bandits,Exploration,Competition,Competition vs Innovation

1 INTRODUCTION

Many modern online platforms simultaneously compete for users as well as learn from the users they manage to attract. This creates a tradeoff between *exploration* and *competition*: firms experiment with potentially sub-optimal options for the sake of gaining information to make better decisions tomorrow, while they need to incentivize consumers to select them over their competitors today. For instance, Google Search and Bing compete for users in the search engine
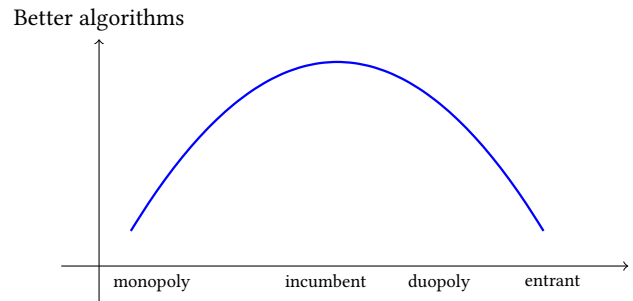
Figure 1: A stylized "inverted-U relationship" between strength of competition and "level of innovation".

market yet at the same time need to experiment with their search and ranking algorithms to learn what works best.

Platforms routinely deploy A/B tests, and are increasingly adopting more sophisticated exploration methodologies based on *multi-armed bandits*, a well-known framework for exploration and making decisions under uncertainty. While deploying "better" learning algorithms for exploration would improve performance, this is not necessarily beneficial under competition, even putting aside the deployment/maintenance costs. In particular, excessive experimentation may hurt platform's reputation and decrease market share in the near term. This would leave the learning algorithms with less users to learn from, which may further degrade platform's performance relative to competitors who keep learning and improving from *their* users, and so forth.

We ask whether algorithms that are traditionally considered "better" in the literature on multi-armed bandits are incentivized to be deployed under competition. We investigate this issue via extensive numerical experiments in a stylized duopoly model. In our model, two firms compete for users and simultaneously learn from them. Each firm commits to a multi-armed bandit algorithm, and *explores* according to this algorithm. Users select between the two firms based on the current reputation score: rewards from the firm's algorithm, averaged over a recent time window. Each firm's objective is to maximize its expected market share (the fraction of users choosing this firm).

We consider a *permanent duopoly* in which both firms start at the same time, as well as *temporary monopoly*: a duopoly with a first-mover. Accordingly, the intensity of competition in the model varies from "permanent monopoly" (just one firm) to "incumbent" (the first-mover in temporary monopoly) to permanent duopoly to "entrant" (late-arriver in temporary monopoly). We find that, even in the permanent duopoly, in equilibrium competition incentivizes firms to play a "greedy algorithm" that does not explicitly explore and even more so if the firm is a late arriver in a market. This algorithm also prevails under monopoly, simply because it tends to be easier to deploy. Whereas the incumbent in the temporary

monopoly is incentivized to deploy a more advanced exploration algorithm that performs better in the long run. As a result, we find that consumer welfare is highest under temporary monopoly and that even increasing the number of firms in the market past a duopoly makes consumers weakly worse off.

Interpreting the adoption of better algorithms as "innovation", our findings can be framed in terms of an "inverted-U relationship" between competition and innovation (see Figure 1). This relationship – too little or too much competition is bad for innovation, but intermediate levels of competition tend to be better – is a familiar theme in the economics literature, dating back to [28]. However, the "inverted-U relationship" is driven by different aspects in our model than the ones in the existing literature in economics. In our case, the barriers for innovations arise entirely from the reputational consequences of exploration in competition, as opposed to the R&D costs (which is the more standard cause in prior work).

The findings on temporary monopoly shed light on the "first-mover advantage" phenomenon in the digital economy. Being first in the market in the digital economy gives firms free data to learn from (a "data advantage") as well as a more definite, and possibly better reputation compared to an entrant (a "reputation advantage").

Beyond the welfare implications, we also investigate the degree to which the incumbent's data advantage serves as a barrier to entry. We run additional experiments that attempt to isolate the impact of the data advantage and reputation advantage. We find that the data advantage from the incumbent's "small head start" can lead to a significant incumbency advantage, especially when the incumbent commits to a more advanced bandit algorithm. Our results show that even a small degree of data asymmetry can be amplified to be a larger market advantage through competition dynamics. This effect is more pronounced when the incumbent commits to a more advanced bandit algorithm since the incumbent does not have to suffer from the reputational consequences of exploration while the entrant does. Our results shed lights on the importance of considering competition dynamics when reasoning about the anti-trust question of whether data can become a significant barrier to entry in an industry.

We also investigate how algorithms' performance "in isolation" (without competition) is predictive of the outcomes under competition. In particular, we find that mean reputation – arguably, the most natural performance measure "in isolation" – is sometimes not a good predictor. We suggest a more refined performance measure, and use it to explain (and frame) some of the competition outcomes.

1.1 Related work

Machine learning. Multi-armed bandits (MAB) is a tractable abstraction for the tradeoff between exploration and *exploitation* (making good near-term decisions based on available information). MAB problems have been studied for many decades, see [14] for background. We consider i.i.d. rewards, a well-studied and well-understood MAB model [6]. We focus on a well-known distinction between "greedy" (exploitation-only) algorithms, "naive" algorithms that separate exploration and exploitation, and "smart" algorithms that combine them. Switching from "greedy" to "naive" to "smart" algorithms involves substantial adoption costs in infrastructure and personnel training [1, 2].

The study of competition vs. exploration has been initiated in [25]. Their model differs from ours in two key respects. First, users do not see any signal about firms' past performance, and instead choose between firms according to the Bayesian-expected reward. Second, they vary the strength of competition using assumptions about (ir)rational consumer behavior, whereas we use early entry. Their results are purely theoretical; their model is amenable to proofs but not to numerical experiments. Their high-level conclusion is an inverted-U relationship between competition and innovation that is similar to ours.

Our setting is also closely related to the "dueling algorithms" framework [20], but this framework considers offline / full feedback scenarios whereas we focus on online machine learning problems.

In "dueling bandits" (e.g., [32, 31]), an algorithm sets up a "duel" between a pair of arms in each round, and only learns which arm has "won". While this setting features a form of competition inside an MAB problem, it is very different from ours.

The interplay between exploration, exploitation and incentives has been studied in other scenarios: incentivizing exploration in a recommendation system, *e.g.,* [22, 17, 15, 24, 12], dynamic auctions (see [11] for background), online ad auctions, *e.g.,* [8, 16, 26, 7], human computation [19, 18, 29], and repeated actions, *e.g.,* [4, 5, 13].

Economics. Our work is related to a longstanding economics literature on competition vs. innovation, *e.g.,* [28, 10, 3]. While this literature focuses on R&D costs of innovation, "reputational costs" seem new and specific to exploration.

Whether data gives competitive advantage has been studied theoretically in [30, 23] and empirically in [9]. While these papers find that small amounts of additional data do not provide significant improvement, they focus on learning in isolation. We ask how additional data helps when firms simultaneously solve a machine learning problem and compete for consumers. Due to the competition dynamics in our model, we find that even a small amount of additional data tends to help a lot.

The first-mover advantage has been well-studied in other settings in economics and marketing, see survey [21].

2 MODEL AND PRELIMINARIES

We consider a game involving two firms and T customers (henceforth, *agents*). The game lasts for T rounds. In each round, a new agent arrives, chooses among the two firms, interacts with the chosen firm, and leaves forever.

Each interaction between a firm and an agent proceeds as follows. There is a set A of K actions, henceforth *arms*, same for both firms and all rounds. The firm chooses an arm, and the agent experiences a numerical reward observed by the firm. Each arm corresponds to a different version of the experience that a firm can provide for an agent, and the reward corresponds to the agent's satisfaction level. The other firm does not observe anything about this interaction, not even the fact that this interaction has happened.

From each firm's perspective, the interactions with agents follow the protocol of the multi-armed bandit problem (MAB). We focus on i.i.d. Bernoulli rewards: the reward of each arm a is drawn from $\{0, 1\}$ independently with expectation $\mu(a)$. The mean rewards $\mu(a)$ are the same for all rounds and both firms, but initially unknown.

Before the game starts, each firm commits to an MAB algorithm, and uses this algorithm to choose its actions. Each algorithm receives a "warm start": additional T_0 agents that arrive before the game starts, and interact with the firm as described above. The warm start serves to vary the initial information the firm has before beginning to compete for users as well as to initialize the agents' perception of the firms. Each firm's objective is to maximize its market share: the fraction of users who chose this firm.

In some of our experiments, one firm is the "incumbent" who enters the market before the other ("late entrant"), and therefore enjoys a *temporary monopoly*. Formally, the incumbent enjoys additional X rounds of the "warm start". We treat X as an exogenous element of the model, and study the consequences for a fixed X.

Agents. For simplicity we suppose that firms compete for agents on a single dimension, quality, that is summarized via the rewards experienced by past consumers. Agents are myopic and non-strategic: they would like to choose among the firms so as to maximize their expected reward (i.e. select the firm with the highest quality), without attempting to influence the firms' learning algorithms or rewards of the future users. Agents are not well-informed: they only receive a rough signal about each firm's performance before they choose a firm, and no other information.

Concretely, each of the two firms has a *reputation score*, and each agent's choice is driven by these two numbers. We posit a version of rational behavior: each agent chooses a firm with a maximal reputation score (breaking ties uniformly). The reputation score is simply a sliding window average: an average reward of the last M agents that chose this firm. This is similar to ratings aggregators on many online platforms such as Yelp, Tripadvisor, etc.

MAB algorithms. We consider three classes of algorithms, ranging from more primitive to more sophisticated:

(1) *Greedy algorithms* that strive to take actions with maximal mean reward, based on the current information.
(2) *Exploration-separating algorithms* that separate exploration and exploitation. The "exploitation" choices strives to maximize mean reward in the next round, and the "exploration" choices do not use the rewards observed so far.
(3) *Adaptive exploration*: algorithms that combine exploration and exploitation, and sway the exploration choices towards more promising alternatives.

We are mainly interested in qualitative differences between the three classes. For concreteness, we fix one algorithm from each class. Our pilot experiments indicate that our findings do not change substantially if other algorithms are chosen. For technical reasons, we consider Bayesian versions initialized with a "fake" prior (*i.e.,* not based on actual knowledge). We consider:

(1) a greedy algorithm that chooses an arm with largest posterior mean reward. We call it "Dynamic Greedy" (because the chosen arm may change over time), DG in short.
(2) an exploration-separated algorithm that in each round, *explores* with probability ε: chooses an arm independently and uniformly at random, and with the remaining probability

exploits according to DG. We call it "dynamic epsilon-greedy", DEG in short.[1]
(3) an adaptive-exploration algorithm called "Thompson Sampling" (TS). In each round, this algorithm updates the posterior distribution for the mean reward of each arm a, draws an independent sample s_a from this distribution, and chooses an arm with the largest s_a.

For ease of comparison, all three algorithms are parameterized with the same fake prior: namely, the mean reward of each arm is drawn independently from a $\mathtt{Beta}(1, 1)$ distribution. Recall that Beta priors with 0-1 rewards form a conjugate family, which allows for simple posterior updates.

Both DEG and TS are classic and well-understood MAB algorithms, see [14, 27] for background. It is well-known that TS is near-optimal in terms of the cumulative rewards, and DEG is very suboptimal, but still much better than DG.[2] In a stylized formula: TS $\gg$ DEG $\gg$ DG as stand-alone MAB algorithms.

MAB instances. We consider instances with $K = 10$ arms. Since we focus on 0-1 rewards, an instance of the MAB problem is specified by the *mean reward vector* $(\mu(a) : a \in A)$. Initially this vector is drawn from some distribution, termed *MAB instance*. We consider three MAB instances:

(1) *Needle-In-Haystack*: one arm (the "needle") is chosen uniformly at random. This arm has mean reward .7, and the remaining ones have mean reward .5.
(2) *Uniform instance*: the mean reward of each arm is drawn independently and uniformly from $[0.25, 0.75]$.
(3) *Heavy-tail instance*: the mean reward of each arm is drawn independently from $\mathtt{Beta}(.6, .6)$ distribution.[3]

We argue that these MAB instances are (somewhat) representative. Consider the "gap" between the best and the second-best arm, an essential parameter in the literature on MAB. The "gap" is fixed in Needle-in-Haystack, spread over a wide spectrum of values under the Uniform instance, and is spread but focused on the large values under the Heavy-Tail instance. We also ran smaller experiments with versions of these instances, and achieved similar qualitative results.

Terminology. Following a standard game-theoretic terminology, algorithm Alg1 *(weakly) dominates* algorithm Alg2 for a given firm if Alg1 provides a larger (or equal) market share than Alg2 at the end of the game. An algorithm is a (weakly) dominant strategy for the firm if it (weakly) dominates all other algorithms. This is for a particular MAB instance and a particular selection of the game parameters.

Simulation details. For each MAB instance we draw $N = 1000$ mean reward vectors independently from the corresponding distribution. We use this same collection of mean reward vectors for all experiments with this MAB instance. For each mean reward vector we draw a table of realized rewards (*realization table*), and use this

---

[1]For our experiments, we fix $\varepsilon = 0.05$. Our pilot experiments showed that different ε did not qualitatively change the results.
[2]Formally, TS achieves regret $\tilde{O}(\sqrt{TK})$ and $O(\frac{1}{\Delta} \log T)$, where Δ is the gap in mean rewards between the best and second-best arms. DEG has regret $\tilde{\Theta}(T^{2/3}K^{1/3})$ in the worst case. And DG can have regret as high as $\Omega(T)$. Deeper discussion of these distinctions is not very relevant to this paper.
[3]This distribution has substantial "tail probabilities".

same table for all experiments on this mean reward vector. This ensures that differences in algorithm performance are not due to noise in the realizations but due to differences in the algorithms in the different experimental settings.

More specifically, the realization table is a 0-1 matrix W with K columns which correspond to arms, and $T + T_{\max}$ rows, which correspond to rounds. Here $T_{\max}$ is the maximal duration of the "warm start" in our experiments, *i.e.,* the maximal value of $X + T_0$. For each arm a, each value $W(\cdot, a)$ is drawn independently from Bernoulli distribution with expectation $\mu(a)$. Then in each experiment, the reward of this arm in round t of the warm start is taken to be $W(t, a)$, and its reward in round t of the game is $W(T_{\max} + t, a)$.

We fix the sliding window size $M = 100$. We found that lower values induced too much random noise in the results, and increasing M further did not make a qualitative difference. Unless otherwise noted, we used $T = 2000$.

The simulations are computationally intensive. An experiment on a particular MAB instance comprised multiple runs of the competition game: N mean reward vectors times 9 pairs of algorithms times three values for the warm start. We used a parallel implementation over a cluster of 12 2.2 GHz CPU cores, with 8 GB RAM per core. Each experiment took about 10 hours.

Consistency. While we experiments with various MAB instances and parameter settings, in the main paper we only report plots for selected, representative experiments. Additional plots will be published in the full version. Unless noted otherwise, our findings are based on and consistent with all these plots.

3 PERFORMANCE IN ISOLATION

We start with a pilot experiment in which we investigate each algorithm's performance "in isolation": in a stand-alone MAB problem without competition. We focus on reputation scores generated by each algorithm. We confirm that algorithms' performance is ordered as we'd expect: TS > DEG > DG for a sufficiently long time horizon. For each algorithm and each MAB instance, we compute the mean reputation score at each round, averaged over all mean reward vectors. We plot the *mean reputation trajectory*: how this score evolves over time. Figure 2 shows such a plot for the Needle In Haystack instance; for other MAB instances the plots are similar. We summarize this finding as follows:

Finding 1. *The mean reputation trajectories are arranged as predicted by prior work:* TS > DEG > DG *for a sufficiently long time horizon.*

The mean reputation trajectory is probably the most natural way to represent an algorithm's performance on a given MAB instance. However, we found that the outcomes of the competition game are better explained with a different "performance-in-isolation" statistic that is more directly connected to the game. Consider the performance of two algorithms, Alg1 and Alg2, "in isolation" on a particular MAB instance. The *relative reputation* of Alg1 (vs. Alg2) at a given time t is the fraction of mean reward vectors/realization tables for which Alg1 has a higher reputation score than Alg2. The intuition is that agent's selection in our model depends only on the comparison between the reputation scores.
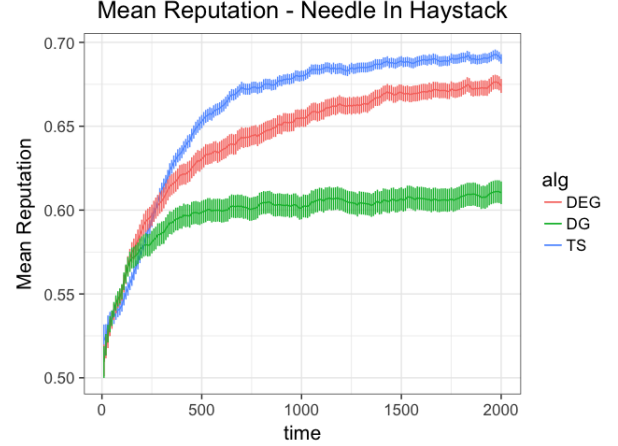

Figure 2: Mean reputation trajectories for Needle-in-Haystack. The shaded area displays 95% confidence intervals.

This angle allows a more nuanced analysis of reputation costs vs. benefits under competition. Figure 3 (top) shows the relative reputation trajectory for TS vs DG for the Uniform instance. The relative reputation is less than $\frac{1}{2}$ in the early rounds, meaning that DG has a higher reputation score in a majority of the simulations, and more than $\frac{1}{2}$ later on. The reason is the exploration in TS leads to worse decisions initially, but allows for better decisions later. The time period when relative reputation vs. DG dips below $\frac{1}{2}$ can be seen as an explanation for the competitive disadvantage of exploration. Such period also exists for the Heavy-tail MAB instance. However, it does not exist for the Needle-in-Haystack instance, see Figure 3 (bottom).[4]

Finding 2. *Exploration can lead to relative reputation vs.* DG *going below* $\frac{1}{2}$ *for some initial time period. This happens for some MAB instances but not for some others.*

Definition 3.1. For a particular MAB algorithm, a time period when relative reputation vs. DG goes below $\frac{1}{2}$ is called *exploration disadvantage period*. An MAB instance is called *exploration-disadvantaged* if such period exists.

Thus, Uniform and Heavy-tail instance are exploration-disadvantaged, but Needle-in-Haystack is not.

4 COMPETITION VS. BETTER ALGORITHMS

Our main experiments are with the duopoly game defined in Section 2. As the "intensity of competition" varies from permanent monopoly to "incumbent" to permanent duopoly to "late entrant", we find a stylized inverted-U relationship as in Figure 1. More formally, we look for equilibria in the duopoly game, where each firm's choices are limited to DG, DEG and TS. We do this for each "intensity level" and each MAB instance, and look for findings that are consistent across MAB instances. For cleaner results, we break ties towards less advanced algorithms (as they tend to have lower

---

[4]We see two explanations for this: TS identifies the best arm faster for the Needle In Haystack instance, and there are no "very bad" arms which make exploration very expensive in the short term.

	Heavy-Tail			Needle-in-Haystack		
	$T_0 = 20$	$T_0 = 250$	$T_0 = 500$	$T_0 = 20$	$T_0 = 250$	$T_0 = 500$
TS vs DG	**0.29** ±0.03 EoG 55 (0)	**0.72** ±0.02 EoG 570 (0)	**0.76** ±0.02 EoG 620 (99)	**0.64** ±0.03 EoG 200 (27)	**0.6** ±0.03 EoG 370 (0)	**0.64** ±0.03 EoG 580 (122)
TS vs DEG	**0.3** ±0.03 EoG 37 (0)	**0.88** ±0.01 EoG 480 (0)	**0.9** ±0.01 EoG 570 (114)	**0.57** ±0.03 EoG 150 (14)	**0.52** ±0.03 EoG 460 (79)	**0.56** ±0.02 EoG 740 (628)
DG vs DEG	**0.62** ±0.03 EoG 410 (7)	**0.6** ±0.02 EoG 790 (762)	**0.57** ±0.03 EoG 730 (608)	**0.46** ±0.03 EoG 340 (129)	**0.42** ±0.02 EoG 650 (408)	**0.42** ±0.02 EoG 690 (467)

Table 1: Permanent duopoly, for Heavy-Tail and Needle-in-Haystack instances. Each cell describes a game between two algorithms, call them Alg1 vs. Alg2, for a particular value of the warm start T_0. Line 1 in the cell is the market share of Alg 1: the average (in bold) and the 95% confidence band. Line 2 specifies the "effective end of game" (EoG): the average and the median (in brackets). The time horizon is $T = 2000$.
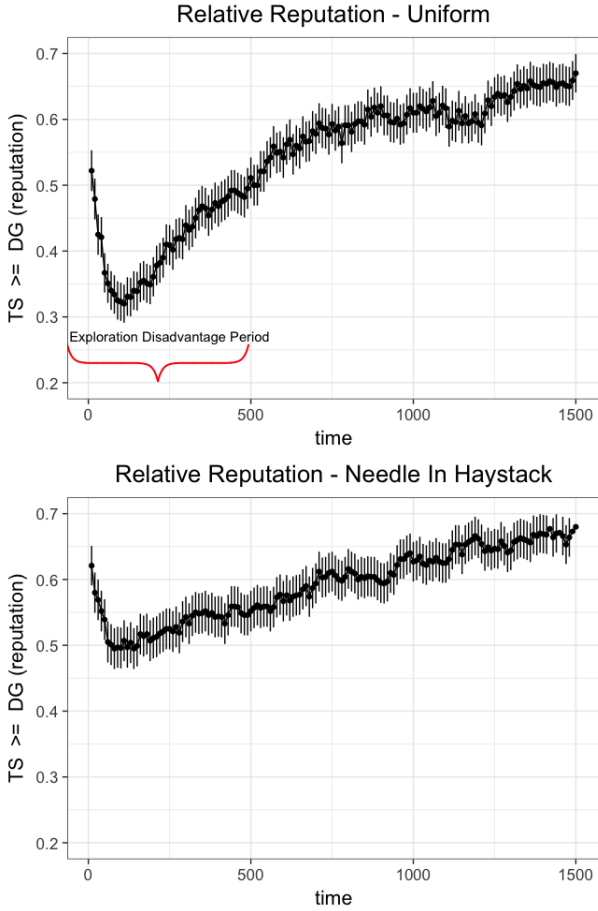

Figure 3: Relative reputation trajectory for TS vs DG, on Uniform instance (top) and Needle-in-Haystack instance (bottom). Shaded area display 95% confidence intervals.

adoption costs [1, 2]). Note that DG is trivially the dominant strategy under permanent monopoly.

Permanent duopoly. The basic scenario is when both firms are competing from round 1. A crucial distinction is whether an MAB instance is exploration-disadvantaged:

Finding 3. *Under permanent duopoly:*

(a) *(DG,DG) is the unique pure-strategy Nash equilibrium for exploration-disadvantaged MAB instances with a sufficiently small "warm start".*

(b) *This is not necessarily the case for MAB instances that are not exploration-disadvantaged. In particular,* TS *is a weakly dominant strategy for Needle-in-Haystack.*

We investigate the firms' market shares when they choose different algorithms (otherwise, by symmetry both firms get half of the agents). We report the market shares for Heavy-Tail and Needle-in-Haystack instances in Table 1 (see the first line in each cell), for a range of values of the warm start T_0. Table 2 reports similarly on the Uniform instance. We find that DG is a weakly dominant strategy for the Heavy-tail and Uniform instances, as long as T_0 is sufficiently small. However, TS is a weakly dominant strategy for the Needle-in-Haystack instance. We find that for a sufficiently small T_0, DG yields more than half the market against TS, but achieves similar market share vs. DG and DEG. By our tie-breaking rule, (DG,DG) is the only pure-strategy equilibrium.

	$T_0 = 20$	$T_0 = 250$	$T_0 = 500$
TS vs DG	**0.46** ±0.03	**0.52** ±0.02	**0.6** ±0.02
TS vs DEG	**0.41** ±0.03	**0.51** ±0.02	**0.55** ±0.02
DG vs DEG	**0.51** ±0.03	**0.48** ±0.02	**0.45** ±0.02

Table 2: Permanent duopoly, for the Uniform MAB instance. Semantics are the same as in Table 1.

We attribute the prevalence of DG on exploration-disadvantaged MAB instances to its prevalence on the initial "exploration disadvantage period", as described in Section 3. Increasing the warm start length T_0 makes this period shorter: indeed, considering relative reputation trajectory in Figure 3 (top), increasing T_0 effectively shifts the starting time point to the right. This is why it helps DG if T_0 is small.

Temporary Monopoly. We turn our attention to the temporary monopoly scenario. Recall that the incumbent firm enters the market and serves as a monopolist until the entrant firm enters at round X. We make X large enough, but still much smaller than the time horizon T. We find that the incumbent is incentivized to choose TS, in a strong sense:

Finding 4. *Under temporary monopoly,* TS *is the dominant strategy for the incumbent. This holds across all MAB instances, if X is large enough.*

The simulation results for the Heavy-Tail MAB instance are reported in Table 3, for a particular $X = 200$. We see that TS is a dominant strategy for the incumbent. Similar tables for the other MAB instances and other values of X are reported in the supplement, with the same conclusion.

	TS	DEG	DG
TS	**0.003**±0.003	**0.083**±0.02	**0.17**±0.02
DEG	**0.045**±0.01	**0.25**±0.02	**0.23**±0.02
DG	**0.12**±0.02	**0.36**±0.03	**0.3**±0.02

Table 3: Temporary monopoly, with $X = 200$ (and $T_0 = 20$), for the Heavy-Tail MAB instance. Each cell describes the duopoly game between the entrant's algorithm (the row) and the incumbent's algorithm (the column). The cell specifies the entrant's market share (fraction of rounds in which it was chosen) for the rounds in which he was present. We give the average (in bold) and the 95% confidence interval. NB: smaller average is better for the incumbent.

DG is a weakly dominant strategy for the entrant, for Heavy-Tail instance in Table 3 and the Uniform instance, but not for the Needle-in-Haystack instance. We attribute this finding to exploration-disadvantaged property of these two MAB instance, for the same reasons as discussed above.

Finding 5. *Under temporary monopoly,* DG *is a weakly dominant strategy for the entrant for exploration-disadvantaged MAB instances.*

Inverted-U relationship. We interpret our findings through the lens of the inverted-U relationship between the "intensity of competition" and the "quality of technology". The lowest level of competition is monopoly, when DG wins out for the trivial reason of tie-breaking. The highest levels are permanent duopoly and "late entrant". We see that DG is incentivized for exploration-disadvantaged MAB instances. In fact, incentives for DG get stronger when the model transitions from permanent duopoly to "late entrant".[5] Finally, the middle level of competition, "incumbent" in the temporary monopoly creates strong incentives for TS. In stylized form, this relationship is captured in Figure 1.

Our intuition for why incumbency creates more incentives for exploration is as follows. During the temporary monopoly period, reputation costs of exploration vanish. Instead, the firm wants to improve its performance as much as possible by the time competition starts. Essentially, the firm only faces a classical explore-exploit tradeoff, and is incentivized to choose algorithms that are best at optimizing this tradeoff.

Death spiral effect. Further, we investigate the "death spiral" effect mentioned in the Introduction. Restated in terms of our model, the effect is that one firm attracts new customers at a lower rate than the other, and falls behind in terms of performance because the other firm has more customers to learn from, and this gets worse

---

[5]For the Heavy-Tail instance, DG goes from a weakly dominant startegy to a strictly dominant one. For the Uniform instance, DG goes from a Nash equilibrium strategy to a weakly dominant one.

over time until (almost) all new customers go to the other firm. With this intuition in mind, we define *effective end of game* (EoG) for a particular mean reward vector and realization table, as the last round t such that the agents at this and previous round choose different firms. Indeed, the game, effectively, ends after this round. We interpret low EoG as a strong evidence of the "death spiral" effect. Focusing on the permanent duopoly scenario, we specify the EoG values in Table 1 (the second line of each cell). We find that the EoG values are indeed small:

Finding 6. *Under permanent duopoly,* EoG *values tend to be much smaller than the time horizon T.*

We also see that the EoG values tend to increase as the warm start T_0 increases. We conjecture this is because larger T_0 tends to be more beneficial for a better algorithm (as it tends to follow a better learning curve). Indeed, we know that the "effective end of game" in this scenario typically occurs when a better algorithm loses, and helping it delays the loss.

Welfare. We study the effects of competition on consumer welfare: the total reward collected by the users over time. Rather than welfare directly, we find it more lucid to consider *market regret*:

$$T \max_a \mu(a) - \sum_{t \in [T]} \mu(a_t),$$

where a_t is the arm chosen by agent t. This is a standard performance measure in the literature on multi-armed bandits. Note that smaller regret means higher welfare.

We assume that both firms play their respective equilibrium strategies for the corresponding competition level. As discussed previously, these are:

- DG in the monopoly,
- DG for both firms in duopoly (Finding 3),
- TS for the incumbent (Finding 4) and DG for the entrant in temporary monopoly (Finding 5).

Figure 4 displays the market regret (averaged over multiple runs) under different levels of competition. Consumers are *better off* in the temporary monopoly case than in the duopoly case. Recall that under temporary monopoly, the incumbent is incentivized to play TS. Moreover, we find that the welfare is close to that of having a single firm for all agents and running TS. We also observe that monopoly and duopoly achieve similar welfare.

Finding 7. *In equilibrium, consumer welfare is (a) highest under temporary monopoly, (b) similar for monopoly and duopoly.*

Finding 7(b) is interesting because, in equilibrium, both firms play DG in both settings, and one might conjecture that the welfare should increase with the number of firms playing DG. Indeed, one run of DG may get stuck on a bad arm. However, two firms independently playing DG are less likely to get stuck simultaneously. If one firm gets stuck and the other does not, then the latter should attract most agents, leading to improved welfare.

To study this phenomenon further, we go beyond the duopoly setting to more than two firms playing DG (and starting at the same time). Figure 5 reports the average welfare across these simulations. Welfare not only does not get better, *but is weakly worse* as we increase the number of firms.

Finding 8. *When all firms deploy* DG, *and start at the same time, welfare is weakly decreasing as the number of firms decreases.*
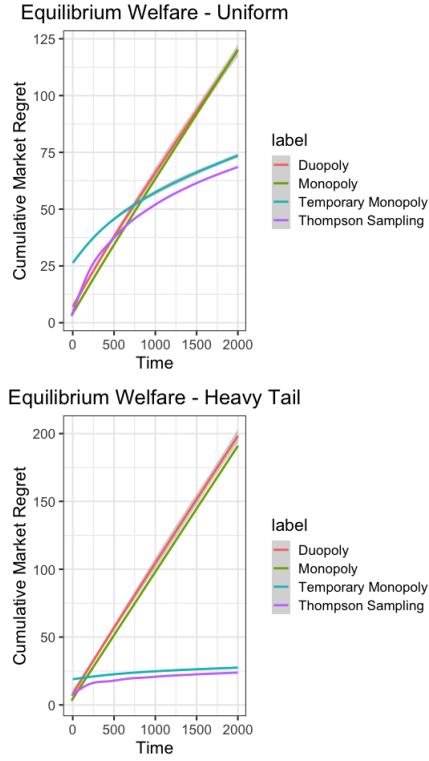
Figure 4: Smoothed welfare plots resulting from equilibrium strategies in the different market structures. Note that welfare at $t = 0$ incorporates the regret incurred during the incumbent and warm start periods.

Figure 5: Average welfare and EoG as we increase #firms playing DG

We track the average EoG in each of the simulations and notice that it *increases* with the number of firms. This observation also runs counter of the intuition that with more firms running DG, one of them is more likely to "get lucky" and take over the market (which would cause EoG to *decrease* with the number of firms).

5 DATA AS A BARRIER TO ENTRY

Under temporary monopoly, the incumbent can explore without incurring immediate reputational costs, and build up a high reputation before the entrant appears. Thus, the early entry gives the incumbent both a *data* advantage and a *reputational* advantage over the entrant. We explore which of the two factors is more significant. Our findings provide a quantitative insight into the role of the classic "first mover advantage" phenomenon in the digital economy.

For a more succinct terminology, recall that the incumbent enjoys an extended warm start of $X + T_0$ rounds. Call the first X of these rounds the *monopoly period* (and the rest is the proper "warm start"). The rounds when both firms are competing for customers are called *competition period*.

We run two additional experiments to isolate the effects of the two advantages mentioned above. The *data-advantage experiment* focuses on the data advantage by, essentially, erasing the reputation advantage. Namely, the data from the monopoly period is not used in the computation of the incumbent's reputation score. Likewise,
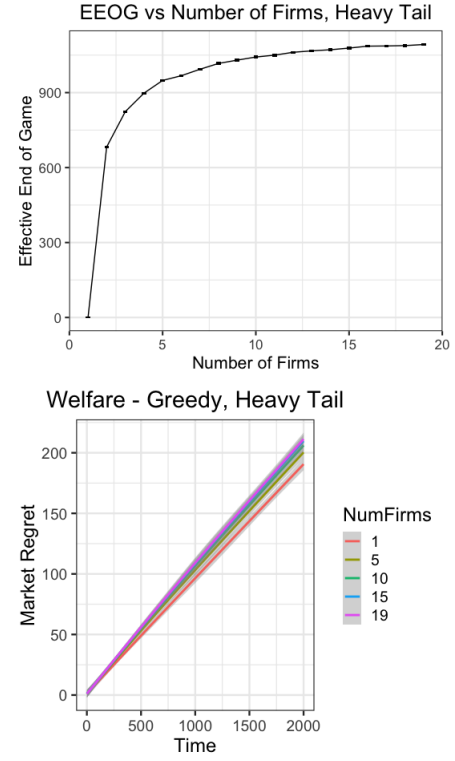
the *reputation-advantage experiment* erases the data advantage and focuses on the reputation advantage: namely, the incumbent's algorithm 'forgets' the data gathered during the monopoly period.

We find that either data or reputational advantage alone gives a substantial boost to the incumbent, compared to permanent duopoly. The results for the Heavy-Tail instance are presented in Table 4, in the same structure as Table 3. For the other two instances, the results are qualitatively similar.

We can quantitatively define the data (resp., reputation) advantage as the incumbent's market share in the competition period in the data-advantage (resp., reputation advantage) experiment, minus the said share under permanent duopoly, for the same pair of algorithms and the same problem instance. In this language, our findings are as follows.

Finding 9.
(a) Data advantage and reputation advantage alone are substantially large, across all algorithms and all MAB instances.
(b) The data advantage is larger than the reputation advantage when the incumbent chooses TS.
(c) The two advantages are similar in magnitude when the incumbent chooses DEG *or* DG.

Our intuition for Finding 9(b) is as follows. Suppose the incumbent switches from DG to TS. This switch allows the incumbent to explore actions more efficiently – collect better data in the same number of rounds – and therefore should benefit the data advantage. However, the same switch increases the reputation cost of

	Reputation advantage			Data advantage		
	TS	DEG	DG	TS	DEG	DG
TS	**0.021**±0.009	**0.16**±0.02	**0.21**±0.02	**0.0096**±0.006	**0.11**±0.02	**0.18**±0.02
DEG	**0.26**±0.03	**0.3**±0.02	**0.26**±0.02	**0.073**±0.01	**0.29**±0.02	**0.25**±0.02
DG	**0.34**±0.03	**0.4**±0.03	**0.33**±0.02	**0.15**±0.02	**0.39**±0.03	**0.33**±0.02

Table 4: Data advantage vs. reputation advantage experiment, on Heavy-Tail MAB instance. Each cell describes the duopoly game between the entrant's algorithm (the row) and the incumbent's algorithm (the column). The cell specifies the entrant's market share for the rounds in which hit was present: the average (in bold) and the 95% confidence interval. NB: smaller average is better for the incumbent.

exploration in the short run, which could weaken the reputation advantage.

6 PERFORMANCE IN ISOLATION, REVISITED

We saw in Section 4 that mean reputation trajectories do not suffice to explain the outcomes under competition. Let us provide more evidence and intuition for this.

Mean reputation trajectories are so natural that one is tempted to conjecture that they determine the outcomes under competition. More specifically:

CONJECTURE 6.1. *If one algorithm's mean reputation trajectory lies above another, perhaps after some initial time interval (e.g., as in Figure 2), then the first algorithm prevails under competition, for a sufficiently large warm start T_0.*

However, we find a more nuanced picture. For example, in Figure 1 we see that DG attains a larger market share than DEG even for large warm starts. We find that this also holds for $K = 3$ arms and longer time horizons, see the supplement for more details. We conclude:

Finding 10. *Conjecture 6.1 is false: mean reputation trajectories do not suffice to explain the outcomes under competition.*

To see what could go wrong with Conjecture 6.1, consider how an algorithm's reputation score is distributed at a particular time. That is, consider the empirical distribution of this score over different mean reward vectors.[6] For concreteness, consider the needle-in-haystack instance at time $t = 500$, plotted in Figure 6. (The other MAB instances lead to a similar intuition.)

We see that the "naive" algorithms DG and DEG have a bi-modal reputation distribution, whereas TS does not. The reason is that for this MAB instance, DG either finds the best arm and sticks to it, or gets stuck on the bad arms. In the former case DG does slightly better than TS, and in the latter case it does substantially worse. However, the mean reputation trajectory may fail to capture this complexity since it simply takes average over different mean reward vectors. This may be inadequate for explaining the outcome of the duopoly game, given that the latter is determined by a simple comparison between the firm's reputation scores.

To further this intuition, consider the difference in reputation scores (*reputation difference*) between TS and DG on a particular mean reward vector. Let's plot the empirical distribution of the

---

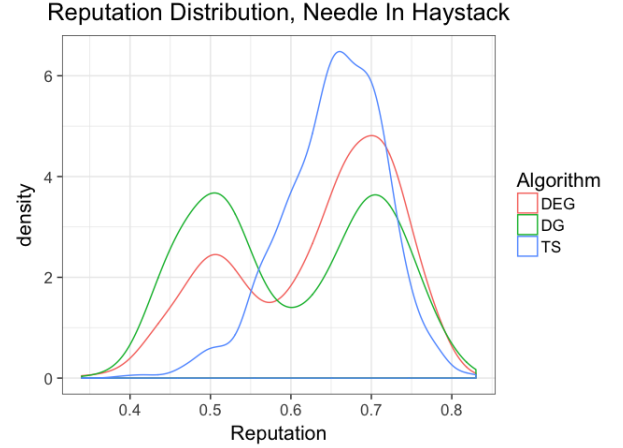[6]Recall that each mean reward vector in our experimental setup comes with one specific realization table.

Figure 6: Distribution of reputation scores for Needle-in-Haystack at $t = 500$ (smoothed using a kernel density estimate)

reputation difference (over the mean reward vectors) at a particular time point. Figure 7 shows such plots for several time points. We observe that the distribution is skewed to the right, precisely due to the fact that DG either does slightly better than TS or does substantially worse. Therefore, the mean is not a good measure of the central tendency, or typical value, of this distribution.

7 CONCLUSION

In this paper we considered a setting where firms simultaneously learn from and compete for users. We showed that competition may not always induce firms to commit to algorithms that engage in exploration resulting in welfare losses for consumers. The primary mechanism behind this was that exploration may have short-term reputational consequences that lead to naive algorithms winning in competition that are bad in the long-run. Allowing one firm to have a head start, or a first-mover advantage, leads to the first-mover having an incentive to play traditionally "good" bandit algorithms and resulting in significantly better welfare for consumers. Finally, we attempted to isolate the effect that the data advantage of an incumbent has on its ability to garner a considerable market share. Contrary to previous studies and discussions which do not consider competition dynamics, we find that the data advantage is substantial in generating significant market power. Additionally we show that this effect is strongest when an incumbent performs efficient
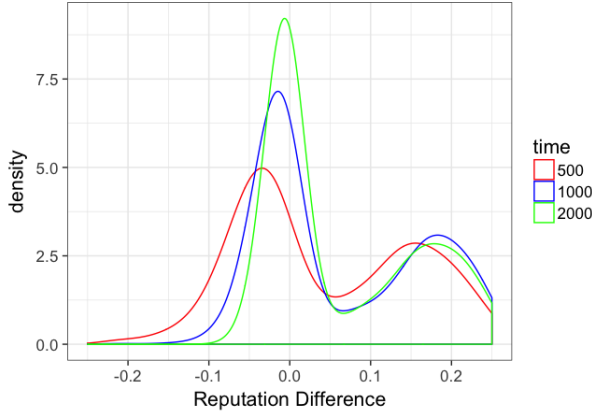
Figure 7: Distribution of reputation difference TS − DG for Needle-in-Haystack (smoothed via a kernel density estimate)

exploration in the incumbency period without the pressure from a competitor.

A   APPENDIX

As stated in Section 2, we only presented plots and tables for selected, representative experiments. We report on all experiments in the full version. In all cases, the plots and tables here are in line with those in the main text, and lead to similar qualitative conclusions. In this appendix, we report on some more experiments.

Temporary Monopoly. We present additional experiments on temporary monopoly from Section 4, across various MAB instances and various values of the incumbent advantage parameter X.

Each experiment is presented as a table with the same semantics as in the main text. Namely, each cell in the table describes the duopoly game between the entrant's algorithm (the row) and the incumbent's algorithm (the column). The cell specifies the entrant's market share (fraction of rounds in which it was chosen) for the rounds in which he was present. We give the average (in bold) and the 95% confidence interval. NB: smaller average is better for the incumbent.

	TS	DEG	DG
TS	**0.17** ±0.02	**0.31** ±0.03	**0.41** ±0.03
DEG	**0.13** ±0.02	**0.26** ±0.02	**0.36** ±0.03
DG	**0.093** ±0.02	**0.23** ±0.02	**0.33** ±0.03

Table 6: Temporary Monopoly: Needle In Haystack, $X = 200$

	TS	DEG	DG
TS	**0.12** ±0.02	**0.16** ±0.02	**0.2** ±0.02
DEG	**0.25** ±0.02	**0.24** ±0.02	**0.29** ±0.02
DG	**0.23** ±0.02	**0.24** ±0.02	**0.29** ±0.02

Table 7: Temporary Monopoly: Uniform, $X = 200$

	TS	DEG	DG
TS	**0.061** ±0.01	**0.12** ±0.02	**0.2** ±0.02
DEG	**0.17** ±0.02	**0.21** ±0.02	**0.29** ±0.02
DG	**0.18** ±0.02	**0.22** ±0.02	**0.29** ±0.02

Table 8: Temporary Monopoly: Uniform, $X = 500$

Reputation vs. Data Advantage We present all experiments on data vs. reputation advantage (Section 5).

Each experiment is presented as a table with the same semantics as in the main text. Namely, each cell in the table describes the duopoly game between the entrant's algorithm (the **row**) and the incumbent's algorithm (the **column**). The cell specifies the entrant's market share for the rounds in which hit was present: the average (in bold) and the 95% confidence interval. NB: smaller average is better for the incumbent.

	TS	DEG	DG
TS	**0.25** ± 0.03	**0.36** ± 0.03	**0.45** ± 0.03
DEG	**0.21** ± 0.02	**0.32** ± 0.03	**0.41** ± 0.03
DG	**0.18** ± 0.02	**0.29** ± 0.03	**0.4** ± 0.03

Table 9: Data Advantage: Needle In Haystack, $X = 200$

	TS	DEG	DG
TS	**0.35** ± 0.03	**0.43** ± 0.03	**0.52** ± 0.03
DEG	**0.26** ± 0.03	**0.36** ± 0.03	**0.43** ± 0.03
DG	**0.19** ± 0.02	**0.3** ± 0.02	**0.36** ± 0.02

Table 10: Reputation Advantage: Needle In Haystack, $X = 200$

	TS	DEG	DG
TS	**0.27** ± 0.03	**0.23** ± 0.02	**0.27** ± 0.02
DEG	**0.4** ± 0.03	**0.3** ± 0.02	**0.32** ± 0.02
DG	**0.36** ± 0.03	**0.29** ± 0.02	**0.3** ± 0.02

Table 11: Reputation Advantage: Uniform, $X = 200$

Heavy Tail		
$T_0 = 20$	$T_0 = 250$	$T_0 = 500$

	Heavy Tail		
	$T_0 = 20$	$T_0 = 250$	$T_0 = 500$
TS vs. DG	**0.4** ±0.02 EoG 770 (0)	**0.59** ±0.01 EoG 2700 (2979.5)	**0.6** ±0.01 EoG 2700 (3018)
TS vs. DEG	**0.46** ±0.02 EoG 830 (0)	**0.73** ±0.01 EoG 2500 (2576.5)	**0.72** ±0.01 EoG 2700 (2862)
DG vs. DEG	**0.61** ±0.01 EoG 1400 (556)	**0.61** ±0.01 EoG 2400 (2538.5)	**0.6** ±0.01 EoG 2400 (2587.5)

**Table 5: Duopoly Experiment: Heavy-Tail,** $K = 3$, $T = 5000$.
Each cell describes a game between two algorithms, call them Alg1 vs. Alg2, for a particular value of the warm start T_0. **Line 1 in the cell is the market share of Alg 1: the average (in bold) and the 95% confidence band. Line 2 specifies the "effective end of game"** (EoG): **the average and the median (in brackets).**

	TS	DEG	DG
TS	**0.2** ± 0.02	**0.22** ± 0.02	**0.27** ± 0.03
DEG	**0.33** ± 0.03	**0.32** ± 0.03	**0.35** ± 0.03
DG	**0.32** ± 0.03	**0.31** ± 0.03	**0.35** ± 0.03

Table 12: Data Advantage: Uniform, $X = 200$

Relative Reputation - Heavy Tail

Mean Reputation vs. Relative Reputation. We present the experiments omitted from Section 6. Namely, experiments on the Heavy-Tail MAB instance with $K = 3$ arms, both for "performance in isolation" and the permanent duopoly game. We find that DEG > DG according to the mean reputation trajectory but that DG > DEG according to the relative reputation trajectory *and* in the competition game. As discussed in Section 6, the same results also hold for $K = 10$ for the warm starts that we consider.

The result of the permanent duopoly experiment for this instance is shown in Table 5. The mean reputation trajectories for algorithms' performance in isolation:

Mean Reputation - Heavy Tail
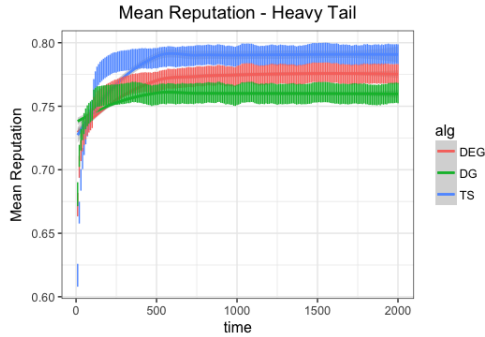
Finally, the relative reputation trajectory of DEG vs. DG:
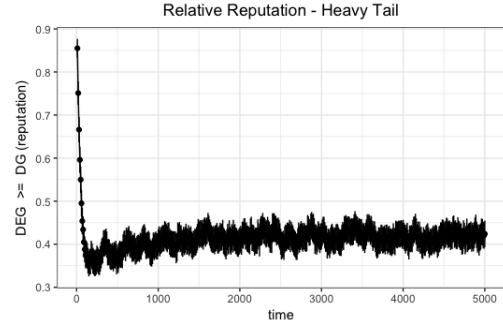
REFERENCES

[1] AGARWAL, A., BIRD, S., COZOWICZ, M., DUDIK, M., LANGFORD, J., LI, L., HOANG, L., MELAMED, D., SEN, S., SCHAPIRE, R., AND SLIVKINS, A. Multiworld testing: A system for experimentation, learning, and decision-making, 2016. A white paper, available at https://github.com/Microsoft/mwt-ds/raw/master/images/MWT-WhitePaper.pdf.
[2] AGARWAL, A., BIRD, S., COZOWICZ, M., HOANG, L., LANGFORD, J., LEE, S., LI, J., MELAMED, D., OSHRI, G., RIBAS, O., SEN, S., AND SLIVKINS, A. Making contextual decisions with low technical debt, 2017. Techical report at arxiv.org/abs/1606.03966.
[3] AGHION, P., BLOOM, N., BLUNDELL, R., GRIFFITH, R., AND HOWITT, P. Competition and innovation: An inverted u relationship. *Quaterly J. of Economics 120*, 2 (2005), 701–728.
[4] AMIN, K., ROSTAMIZADEH, A., AND SYED, U. Learning prices for repeated auctions with strategic buyers. In *26th NIPS* (2013), pp. 1169–1177.
[5] AMIN, K., ROSTAMIZADEH, A., AND SYED, U. Repeated contextual auctions with strategic buyers. In *27th NIPS* (2014), pp. 622–630.
[6] AUER, P., CESA-BIANCHI, N., AND FISCHER, P. Finite-time analysis of the multi-armed bandit problem. *Machine Learning 47*, 2-3 (2002), 235–256.
[7] BABAIOFF, M., KLEINBERG, R., AND SLIVKINS, A. Truthful mechanisms with implicit payment computation. *J. ACM 62*, 2 (2015), 10. Subsumes the conference papers in *ACM EC 2010* and *ACM EC 2013*.
[8] BABAIOFF, M., SHARMA, Y., AND SLIVKINS, A. Characterizing truthful multi-armed bandit mechanisms. *SIAM J. on Computing 43*, 1 (2014), 194–230. Preliminary version in *10th ACM EC*, 2009.
[9] BAJARI, P., CHERNOZHUKOV, V., HORTAÇSU, A., AND SUZUKI, J. The impact of big data on firm performance: An empirical investigation. Tech. rep., National Bureau of Economic Research, 2018.
[10] BARRO, R. J., AND SALA-I MARTIN, X. Economic growth: Mit press. *Cambridge, Massachusettes* (2004).
[11] BERGEMANN, D., AND SAID, M. Dynamic auctions: A survey. In *Wiley Encyclopedia of Operations Research and Management Science, Vol. 2*. Wiley: New York, 2011, pp. 1511–1522.
[12] BIMPIKIS, K., PAPANASTASIOU, Y., AND SAVVA, N. Crowdsourcing exploration. *Management Science* (2018). Forthcoming. Published online as *Articles in Advance* in April 2017.
[13] BRAVERMAN, M., MAO, J., SCHNEIDER, J., AND WEINBERG, M. Selling to a no-regret buyer. In *ACM EC* (2018), pp. 523–538.
[14] BUBECK, S., AND CESA-BIANCHI, N. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning 5*, 1 (2012).

[15] CHE, Y.-K., AND HÖRNER, J. Optimal design for social learning. *Quarterly Journal of Economics* (2018). Forthcoming. First published draft: 2013.

[16] DEVANUR, N., AND KAKADE, S. M. The price of truthfulness for pay-per-click auctions. In *10th ACM EC* (2009), pp. 99–106.

[17] FRAZIER, P., KEMPE, D., KLEINBERG, J. M., AND KLEINBERG, R. Incentivizing exploration. In *ACM EC* (2014), pp. 5–22.

[18] GHOSH, A., AND HUMMEL, P. Learning and incentives in user-generated content: multi-armed bandits with endogenous arms. In *ITCS* (2013), pp. 233–246.

[19] HO, C.-J., SLIVKINS, A., AND VAUGHAN, J. W. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *J. of Artificial Intelligence Research 55* (2016), 317–359. Preliminary version appeared in *ACM EC 2014*.

[20] IMMORLICA, N., KALAI, A. T., LUCIER, B., MOITRA, A., POSTLEWAITE, A., AND TENNENHOLTZ, M. Dueling algorithms. In *43rd ACM STOC* (2011), pp. 215–224.

[21] KERIN, R. A., VARADARAJAN, P. R., AND PETERSON, R. A. First-mover advantage: A synthesis, conceptual framework, and research propositions. *The Journal of Marketing* (1992), 33–52.

[22] KREMER, I., MANSOUR, Y., AND PERRY, M. Implementing the âĂIJwisdom of the crowdâĂİ. *J. of Political Economy 122* (2014), 988–1012. Preliminary version in *ACM EC 2014*.

[23] LAMBRECHT, A., AND TUCKER, C. E. Can big data protect a firm from competition?

[24] MANSOUR, Y., SLIVKINS, A., AND SYRGKANIS, V. Bayesian incentive-compatible bandit exploration. In *15th ACM EC* (2015).

[25] MANSOUR, Y., SLIVKINS, A., AND WU, S. Competing bandits: Learning under competition. In *9th ITCS* (2018).

[26] NAZERZADEH, H., SABERI, A., AND VOHRA, R. Dynamic cost-per-action mechanisms and applications to online advertising. In *17th WWW* (2008).

[27] RUSSO, D., ROY, B. V., KAZEROUNI, A., OSBAND, I., AND WEN, Z. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning 11*, 1 (2018), 1–96.

[28] SCHUMPETER, J. *Capitalism, Socialism and Democracy*. Harper & Brothers, 1942.

[29] SINGLA, A., AND KRAUSE, A. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *22nd WWW* (2013), pp. 1167–1178.

[30] VARIAN, H. Artificial intelligence, economics, and industrial organization. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 2018.

[31] YUE, Y., BRODER, J., KLEINBERG, R., AND JOACHIMS, T. The k-armed dueling bandits problem. *J. Comput. Syst. Sci. 78*, 5 (2012), 1538–1556. Preliminary version in COLT 2009.

[32] YUE, Y., AND JOACHIMS, T. Interactively optimizing information retrieval systems as a dueling bandits problem. In *26th ICML* (2009), pp. 1201–1208.