

# Learning and Reputation under Monopoly and Duopoly: A Competing Bandits Approach

Guy Aridor<sup>1</sup>, Kevin Liu<sup>2</sup>, Aleksandrs Slivkins<sup>3</sup>, Zhiwei Steven Wu<sup>4</sup>

<sup>1</sup>Columbia University, Department of Economics

<sup>2</sup>Columbia University, Department of Computer Science

<sup>3</sup>Microsoft Research, New York, NY

<sup>4</sup>University of Minnesota - Twin Cities, Department of Computer Science

## Abstract

We empirically study the interplay between *exploration* and *competition* - how systems that learn from interactions with users tradeoff between making potentially suboptimal choices in order to acquire new information and the reputational consequences of this exploration which potentially has adverse competitive effects. Our model considers competition for myopic users between two firms deploying multi-armed bandit algorithms that face the same underlying multi-armed bandit instance. The users select between the firms according to a reputation score, which is a function of the rewards past users have experienced from this firm.

We ask whether better algorithms are incentivized under varying degrees of competition. The environments we consider are a monopoly, a duopoly with one firm serving as an early entrant (a temporary monopoly), and a duopoly. We show that, under certain conditions, monopoly and duopoly do not incentivize the adoption of better learning algorithms due to the reputational costs of exploration, but that a temporary monopoly may incentivize the adoption of better learning algorithms. Additionally, we interpret our results as providing an alternative intuition behind the classic first-mover advantage which gives the incumbent firm both a data advantage and a reputational advantage. Finally, we ask whether in this setting the data advantage or the reputational advantage of the early entrant serves as stronger barriers to entry.

## Introduction

Many modern online platforms simultaneously compete for users as well as learn from the users they manage to attract. For instance, Google Search and Bing compete for users in the search engine market yet at the same time need to experiment with their search algorithms to learn what algorithms work the best. This creates a tradeoff between *exploration* and *competition* since firms need to experiment with potentially sub-optimal options for the benefit of gaining information to make better decisions tomorrow while at the same time firms need to incentivize consumers to select them over their competitors today.

From a social welfare perspective it is better for firms to adopt "better" learning algorithms and our main high-level question is asking under what levels of competition are firms incentivized to adopt "better" learning algorithms. Our primary contribution is that in our model, even in the case of a duopoly, firms are not incentivized to adopt better learning algorithms but that allowing one firm to be a temporary monopolist will incentivize this firm to play a better learning algorithm.

Unlike in the classic models in economics regarding the interplay between competition and innovation described in Barro and Sala-i Martin (2004), here we have no explicit R & D cost. Rather, firms are not incentivized to experiment under duopoly because exploration has an implicit relative *reputational* cost compared to the greedy alternative and, in our model, consumers only care about the reputation of a firm. Thus, competition disincentivizes exploration by firms precisely because competition forces firms to take actions in order to incentivize consumers to select them over their competitors and, as a result, by allowing a firm to be a temporary monopolist the firm can experiment without worrying about the reputational consequences. Our findings give an alternative intuition to the empirically documented inverse-U relationship between competition and innovation discussed in Aghion et al. (2005).

**Related Work** A major underlying component of our model is that firms face a multi-armed bandit (MAB) problem. Multi-armed bandits (MAB) are a tractable abstraction for the tradeoff between exploration and exploitation. MAB problems have been studied for many decades (see Bubeck, Cesa-Bianchi, and others (2012) for an overview). Most relevant to this paper is the thread that focuses on designing "smart" and tractable algorithms that combine exploration and exploitation and "naive" algorithms that separate exploration and exploitation (see Slivkins (2017)).

The three-way tradeoff between exploration, exploitation, and incentives has been studied in several other settings, the most relevant to this paper being Che and Hörner 2017, Kremer, Mansour, and Perry 2014, Mansour, Slivkins, and Syrgkanis 2015. The strategic experimentation literature in economics, such as Bolton and Harris 1999 and Keller, Rady, and Cripps 2005, stud-

ies models with self-interested agents jointly performing exploration whereas in this paper the firms cannot observe the actions or the payoffs of the other firms and exploration is coordinated by the consumers.

The relationship between competition and innovation has been heavily studied in industrial organization (Tirole, 1988) and endogenous growth theory (Aghion et al., 2005; Barro and Sala-i Martin, 2004), dating back to Schumpeter (2010).

Our setting is also closely related to the "dueling algorithms" framework described in Immorlica et al. (2011), but this framework considers offline scenarios whereas we focus on online learning problems.

The most closely related work to this paper is Mansour, Slivkins, and Wu 2018. Our motivating questions are the same as Mansour, Slivkins, and Wu (2018) though while Mansour, Slivkins, and Wu (2018) use the rationality of consumers as their primary knob of competitiveness, we consider differences in the number of firms in the market as our primary knob. Additionally, while in Mansour, Slivkins, and Wu (2018) agents select firms based on which firm they expect to have a larger Bayesian expected reward, in our model the agents select firms based on a frequentist reputation score that is derived from the signals sent from past agents. This allows the model to become tractable for the purposes of evaluating it via simulation and this is the primary method of analysis that we employ.

## Model

**Overview** There are two firms and  $T + 2k$  agents where  $k$  is the warm start, or the agents that each firm gets for free at the beginning of the game and  $T$  is the number of rounds in the game. The timing of events is as follows:

1. At  $t = 0$ , the firms simultaneously commit to following a learning algorithm from a set of algorithms  $\mathcal{A}$
2. Still at  $t = 0$  we suppose that each firm gets  $k$  agents as a "warm" start. The algorithm that the firm commits to makes  $k$  rounds of progress and uses this information to initialize its information set. Additionally, the reputation score used by the agents is initialized using the rewards from these  $k$  rounds.
3. A new agent arrives each round (and lives for only one round), starting at  $t = 1$ , and chooses the firm with the higher reputation at the time.
4. The firm that is chosen selects an action  $a_t \in A$  from a set of actions that is fixed across firms and rounds.
5. Both the agent and the firm observe the reward  $r_t \in [0, 1]$  from the action. The agent reports this reward to the firm and the future agents and the reputation score for the chosen firm is updated.
6. Repeat 2-4 for  $T$  rounds.

Generally, the rewards are independent and identically distributed with a common prior. For computational tractability we restrict our focus to Bernoulli-distributed rewards with Beta priors. Each firm faces a multi-armed

bandit problem with no initial information<sup>1</sup>. We assume that the firms commit to a multi-armed bandit learning algorithm at the start of the world and that there are no informational spillovers from their competitors so that they can only learn from the agents that select them.

**Agents** We suppose that agents are homogenous, myopic, and non-strategic. The utility function for the agents is simply to maximize their reward in the one period in which they are alive. We suppose that agents do not attempt to manipulate the strategy of the firm, do not take the strategies of the firm into account when choosing between the firms, and do not condition their behavior on calendar time. In our model, each agent uses the average reward of past agents as a proxy for their expected utility. For simplicity, the reputation score,  $R_{jt}$  is defined as a sliding window average<sup>2</sup>.

$$R_{jt} = \frac{1}{M} \sum_{i=1}^M r_{t_j-i}$$

Note  $t_j$  is the *local* time of the firm, not the global time, so the reputation score is the sliding window average of the last  $M$  times that firm  $j$  was selected by the agents. The warm start of  $k$  rounds allows this reputation score to be well-defined once the "competition game" begins. In our model the agents deterministically choose the firm with the higher reputation score and ties are broken uniformly at random.

**Firms** We suppose that the firms simply care about maximizing their expected market share (i.e. maximizing the number of  $T$  agents who select them). We model the "competition game" between firms as a simultaneous move game where both firms commit to a learning algorithm at  $t = 0$ .

**MAB algorithms** We suppose that firms commit to a learning algorithm from a fixed set of algorithms  $\mathcal{A}$ . We partition the set of possible learning algorithms into three different types of learning algorithms and restrict  $\mathcal{A}$  to contain a representative each algorithm from each class:

1. "Smart" algorithms that engage in adaptive exploration and combine exploration and exploitation. We consider *Thompson Sampling* (from hereon *TS*) from this class, which, in a given period, will pull an arm according to the probability that that arm is "optimal" in the sense of having the highest mean reward (Agrawal and Goyal, 2012).
2. "Naive" algorithms that engage in non-adaptive exploration algorithms and separate exploration and exploitation. We consider *Dynamic  $\epsilon$ -greedy* (from

<sup>1</sup>For algorithms that require a prior to operate, such as Thompson Sampling, we use a "fake" prior of  $Beta(\alpha = 1, \beta = 1)$

<sup>2</sup>Another natural formulation would be to have exponential discounting of the past rewards. We believe that the results should be qualitatively similar to those presented here as in both formulations the more recent past matters more than the distant past.

hereon *DEG*) from this class, which, in a given period, pulls the arms with the highest posterior mean for  $1 - \epsilon$  probability and selects a random arm with  $\epsilon$  probability. For our experiments we keep  $\epsilon = 0.05$  fixed.

3. Greedy / myopic algorithms that engage in no purposeful exploration and take the best short-sighted action. We consider *DynamicGreedy* (from hereon *DG*) from this class, which, in a given period, pulls the arm with the highest posterior mean.

In the standard multi-armed bandit problem it is known that  $TS > DEG > DG$  in terms of maximizing cumulative reward over a sufficiently large time horizon. The primary question that we want to understand is when, in competition, are the firms incentivized to adopt the "better" algorithms?

**Incumbent** In some of our experiments we modify our model so that one firm enters the market  $X$  number of rounds before the other. We refer to the firm that enters before as the "incumbent" and the firm that enters after  $X$  rounds as the "entrant." In the  $X$  rounds before the "entrant" enters the incumbent is a temporary monopolist as the agents that arrive in these  $X$  rounds are forced to select the incumbent since it is the only firm in the market and there is no outside option. We treat  $X$  as being an exogenous element of the model and study the consequences for a fixed  $X$ . We have that both the incumbent and entrant commit to a learning algorithm *before* either firm receives any agent. After the  $X$  rounds, both firms still receive the  $k$  warm start agents.

## Simulation Details

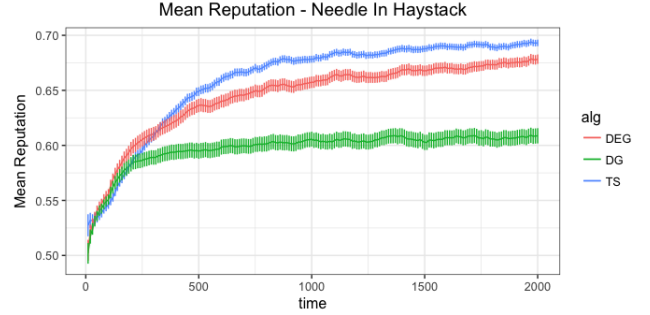
We evaluate the consequences of our model via simulation.

**Bandit Priors** We look at bandit instances drawn from three different bandit "priors" that are different types of learning problems. Recall that we only consider Bernoulli-distributed arms and so we draw the true means of the arms from these priors. For these experiments we fix the number of arms we consider,  $K = 10$ .

1. Needle In Haystack -  $K - 1$  arms with mean 0.5 and 1 arm with 0.7.
2. Uniform - the  $K$  arms have means drawn uniformly at random from  $[0.25, 0.75]$
3. "Heavy Tail" - the  $K$  arms have means drawn from  $Beta(\alpha = 0.6, \beta = 0.6)$ . With this prior it was likely to have means at the "extremes" or means that were close to 0 as well as means that were close to 1.

**Simulation of Competition Game** Unless otherwise noted, all of the reported results utilize the same set of randomly drawn bandit instances and realizations from the prior. Namely, for each bandit prior we draw  $N = 1000$  bandit instances. For each of these instances we run simulations of our model for varying values of  $k$  and  $X$ . We take the maximum values of  $k$  and  $X$  for the

Figure 1: Mean Reputation Trajectories in Isolation



The plots contain the average reputation over 1000 runs for a memory size of 100 where, for a given  $t$ , we record the reputation of a given algorithm on a given instance and then average this value across all the runs. The shaded area display 95% confidence intervals.

simulations we run,  $k_{max}$  and  $X_{max}$  respectively, and compute a realization table of dimension  $(T + k_{max} + X_{max}) \times K$ .

This realization table, as well as fixing the random seed for the same bandit instance and realization table across experiments, 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. In the competition game we draw from the  $T \times K$  portion of the table, so that if two different algorithms picks arm  $a$  at time  $t$ , they get the same  $[a, t]$  realization in the table. This setup also ensures that in the warm start period, increasing the warm start from  $k$  to  $k + 10$  results in the same behavior in the first  $k$  rounds.

For the simulations we fix the sliding window size  $M = 100$ . Low values of  $M$  induced too much random noise into the results and we found that increasing  $M$  to be larger than 100 did not make a substantial qualitative difference so we fix this value.

## Performance in Isolation

We evaluate the performance of the different learning algorithms on the bandit priors we consider in isolation. There are two reasons why this is important to consider. First, we want to confirm that the reputation ordering of the algorithms is what we would expect according to the multi-armed bandit literature where, according to the mean reputation,  $TS > DEG > DG$  for sufficiently large  $t$ . Second, we want to better understand what statistics of the priors we can look at in isolation in order to help us predict and understand what to expect in the competition game.

Figure 1 shows that the mean reputation ordering is as we would expect for the Needle In Haystack prior<sup>3</sup>.

<sup>3</sup>The mean reputation ordering that we expect also holds for the other two priors and the results for those can be found in the supplementary material.

**Finding 1.** *The mean performance of an algorithm is not a sufficient statistic for understanding the performance of an algorithm in competition. Looking at the distribution of reputation difference between algorithms we find that it tend is skewed to the right. Additionally we provide a counterexample where the mean performance does not predict the result of the competition game.*

Is looking at the mean performance of between two algorithms sufficient for understanding what happens in competition? To answer this, we first look at the entire reputation distribution at  $t = 500$  in Figure 2. We see that the “naive” algorithms *DG* and *DEG* have a bi-modal reputation distribution whereas *TS* does not. The intuition for this is that, for Needle In Haystack, *DG* either finds the best arm or not. If it does, then since it engages in no purposeful exploration it will do better than any algorithm that engages in purposeful exploration over sufficiently many rounds. However, if it does not then it will get stuck on a bad arm and lose to *TS* or *DEG*. In these cases its reputation may be substantially worse but for the competition game the relative comparison between them is all that matters <sup>4</sup>.

To better capture the comparison between two algorithms we explicitly plot the distribution of the reputation difference between *TS* and *DG*. This figure seems to confirm the intuition noted previously since the reputation distribution has its largest mass around the point just below 0 but it is skewed to the right even as  $t$  gets large. Given the skewedness of the distribution, the mean is not a representative statistic of the entire reputation difference distribution. There are alternative statistics that we could consider such as the median or numerically calculating  $\Pr(\text{reputation}(TS) - \text{reputation}(DG)) > 0$  using the estimated density, but instead we define a simple and natural statistic to help us reason about what happens under competition, the *relative reputation proportion*.

**Definition 1.** *Relative Reputation Proportion* - the proportion of simulations in which algorithm  $A$  had at least as high of a reputation as algorithm  $B$  for a fixed time  $t$

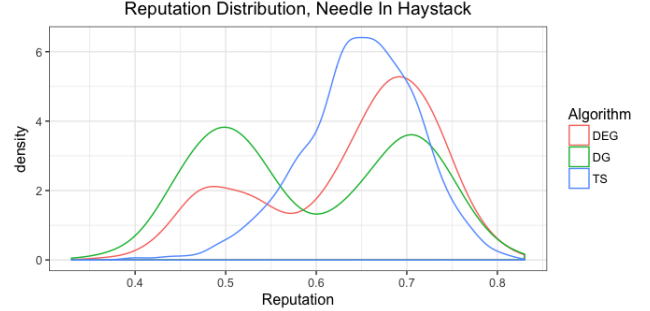
**Finding 2.** *Purposeful exploration can lead to relative reputational costs compared to the greedy alternative and this leads to *TS* doing worse than *DG* for small time horizons. We observe this under the Uniform and Heavy Tail priors. However, it is not observed under the Needle In Haystack prior*

The relative reputation statistic corresponds to running the bandit algorithms in isolation on the same instance and with the same realizations for  $t$  rounds and then calculating the fraction of simulations at which an agent would select a firm playing  $A$  over a firm playing  $B$  at time  $t$  <sup>5</sup>.

Figure 4 shows plots of the relative reputation proportion for *TS* vs *DG* on the Uniform and Needle In

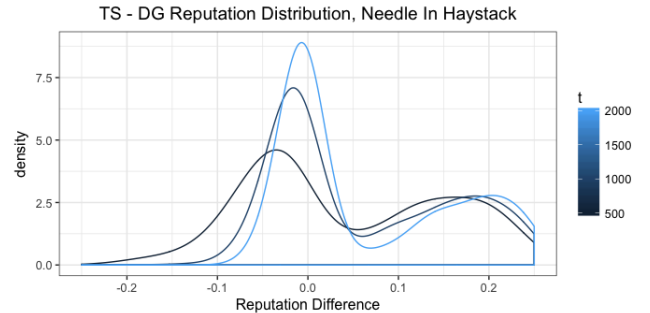
<sup>4</sup>This holds for our model and the decision rule of the agents, though the absolute difference may matter if, for instance, we consider the SoftMax decision rule in Mansour,

Figure 2: Reputation Distribution



The plots contain a kernel density estimate of the reputation distribution at  $t = 500$

Figure 3: Reputation Difference Distribution



The plots contain a kernel density estimate of the difference in reputation between *TS* and *DG* across  $t$

Haystack prior. For the Uniform prior we see that, in the early rounds,  $DG > TS$  for the majority of the simulations but that, eventually,  $TS > DG$ . The intuition behind this is that, especially since the firms start with no substantive initial information,  $TS$  does purposeful exploration in the early rounds in order to acquire information. However, eventually the information acquired from purposeful exploration in the early rounds allows  $TS$  to make better decisions and achieve a higher reputation, especially when the instance is “hard enough” so that  $DG$  cannot trivially find the best arm. The early exploration leads to what we define as a *relative reputation cost* and the eventual gain in reputation leads to what we define as a *relative reputation benefit*.

**Definition 2.** *Relative reputation cost and benefit* - the relative reputation loss an algorithm incurs from purposeful exploration compared to the greedy alternative. Here we treat it as an empirical definition based on the relative reputation proportion where the “cost” regime is when the proportion  $< 0.5$  and the “benefit” regime is when the proportion is  $> 0.5$  for the better algorithm. We call the priors where there is an early costly period followed by a later benefit period relative reputation costly priors.

Is exploration always costly? Figure 4 also shows that for the Needle In Haystack prior,  $TS$  always does relatively better than  $DG$ . There are two contributing factors to this. First,  $TS$  identifies the best arm faster in the Needle In Haystack prior than the Uniform prior so that there is a shorter time horizon where  $TS$  needs to engage in purposeful exploration. Second, in the Needle In Haystack prior there are no “bad” arms as there may be in the Uniform prior since by construction all the arms except one in Needle In Haystack are the same. Thus, when  $TS$  pulls a sub-optimal arm relative to its current information, the expected reward is the same as the greedy option that has not identified the best arm. However, with the Uniform prior, it is possible that the sub-optimal arm that is pulled has substantially lower expected reward relative to the greedy option. Thus, only the Uniform and Heavy Tail priors are relative reputation costly priors.

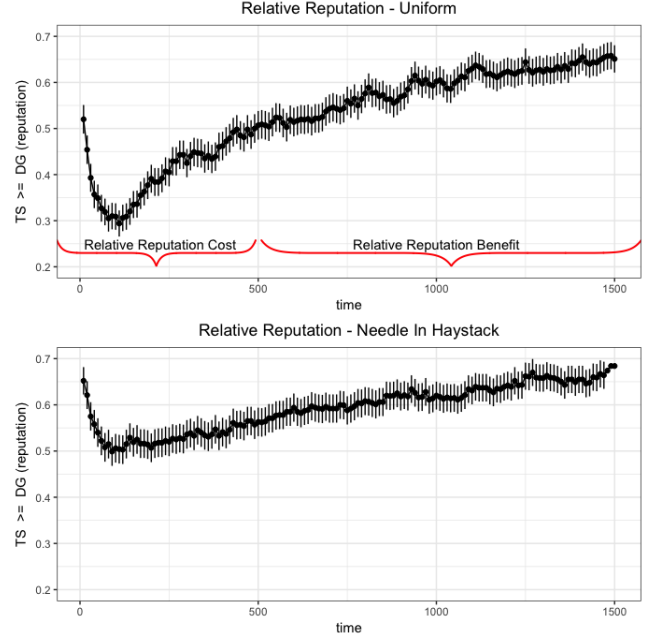
## Competition vs Better Algorithms: Inverted-U

In this section we document our main results that can be summarized via the empirically documented inverted-U

Slivkins, and Wu (2018)

<sup>5</sup>As further motivation for this statistic, one may be interested in if there is ever a case where, for sufficiently large  $t$ , we observe that  $DEG > DG$  or  $TS > DG$  according to the mean reputation but  $DG > DEG$  or  $DG > TS$  according to the relative reputation proportion. In the supplementary material we have results showing that, for Heavy Tail prior with  $K = 3$  we have that  $DEG > DG$  according to the mean reputation but  $DG > DEG$  according to the relative reputation proportion. Additionally, these results carry over to the competition game

Figure 4: Relative Reputation Plots



The plots contain the average reputation over 1000 runs for a memory size of 100 where, for a given  $t$ , we record the reputation of both of the algorithms on a given instance and then calculate the proportion of runs where  $TS \geq DG$ . The shaded area display 95% confidence intervals.

relationship between competition and innovation described in Aghion et al. (2005). We consider three separate settings for our model: permanent monopoly, temporary monopoly (incumbent vs entrant), and permanent duopoly.

**Finding 3.** *Under the priors that are relative reputation costly and with low warm start, the optimal strategies in the competition game are:*

**Permanent Monopoly - DG**  
**Temporary Monopoly - TS**  
**Permanent Duopoly - DG**

*The conditions on TS being dominant under the temporary monopoly are that the incumbent is a temporary monopoly for sufficiently many periods.*

We interpret the move from permanent monopoly to temporary monopoly and temporary monopoly to permanent duopoly as increases in competition. We believe this is reasonable and consistent with the empirical results which define increases in competition as increases in a measure of market power (1 - the Lerner Index). In our model we have no prices and no costs and so we interpret increases in “market power” as decreases in the number of rounds where a firm has to compete for agents.

We utilize the same instances and realizations as in the section 4 and simulate the model described in section 2. We initially take the strategies of the firms as exogenous (chosen from  $\mathcal{A}$ ) and simulate the model given these

strategies in order to determine the expected payoffs associated with each pair of algorithms. Unless otherwise noted, all the results are reported at  $t = 2000$ .

**Permanent Monopoly** Under permanent monopoly only a single firm exists in the market for every period. Since, in our formulation, the firm only gets utility from a larger market share it is indifferent between each algorithm in  $\mathcal{A}$  since, regardless of the algorithm it deploys, it will get the entire market. If we suppose that deploying the "better" algorithms has even an  $\epsilon > 0$  cost, then the firm would choose to deploy  $DG$ .

**Permanent Duopoly** Permanent duopoly corresponds to the model described in section 2 where  $X = 0$  so that both firms enter the market simultaneously. What algorithms are firms incentivized to deploy? We report the average market share taken over the  $N$  simulations for a set of exogenous values of  $k$ . Additionally, we report the mean and median of an additional quantity which we define as the effective end of game.

**Definition 3.** *Effective End of Game (EEOG)* - the last round,  $t$ , in a simulation where the agent alive at  $t - 1$  and the agent alive at  $t$  choose different firms.

Since both firms enter at the same time the game is symmetric and we only need to compute the payoffs of three pairs of strategies. We do not report them in the figures, but when both firms play the same algorithm the expected market share is 50/50.

**Finding 4.** *For low warm starts, when exploration is relative reputation costly  $DG > TS$  in competition, whereas when exploration is not relative reputation costly  $TS > DG$ . For sufficiently large warm starts,  $TS$  is dominant across all priors.*

Table 1 shows the results under the Uniform prior. For the low values of the warm start we can see that  $DG > TS$ , but that for larger warm starts,  $TS > DG$ <sup>6</sup>. Looking at the relative reputation plots in Figure 4, we can interpret fixing a warm start  $k$  as fixing the starting point on the relative reputation plots. The proportion of first rounds in the competition game that will go to a firm playing alg  $A$  over a firm playing alg  $B$  will correspond to the relative reputation proportion at time  $k$ . The intuition for the result in competition aligns with the fact that, for Uniform and Heavy Tail,  $TS$  is relative reputation costly compared to  $DG$  and so, for low warm starts,  $TS$  has not yet recovered from the relative reputation loss due to purposeful exploration and so  $DG$  beats  $TS$ . However, if we move the warm start sufficiently high so that  $TS$  moves to the relative reputation benefit regime,  $TS > DG$ .

However, we do not always see that  $DG > TS$  under some warm starts. Table 2 shows the results for the Needle In Haystack prior where we see that  $TS > DG$ . This is precisely due to the fact that, under the Needle In Haystack,  $TS$  is not relative reputation costly compared to  $DG$  and thus we see that  $TS > DG$ .

<sup>6</sup>The same results hold for the Heavy Tail prior and are reported in the supplemental material

Combining these results with the fact that the EEOG values are relatively low across all warm starts<sup>7</sup>, these observations seem to imply that understanding the performance of different algorithms in competition it is important to look at their performance for relatively small samples instead of asymptotically.

Table 1: Duopoly Experiment Uniform

	k = 20	k = 250	k = 500
TS vs DG	<b>0.46</b> $\pm 0.03$	<b>0.52</b> $\pm 0.02$	<b>0.6</b> $\pm 0.02$
	eeog	eeog	eeog
	avg: 230 med: 0	avg: 800 med: 754	avg: 910 med: 906.5
TS vs DEG	<b>0.41</b> $\pm 0.03$	<b>0.51</b> $\pm 0.02$	<b>0.55</b> $\pm 0.02$
	eeog	eeog	eeog
	avg: 180 med: 0	avg: 810 med: 734	avg: 970 med: 987
DG vs DEG	<b>0.51</b> $\pm 0.03$	<b>0.48</b> $\pm 0.02$	<b>0.45</b> $\pm 0.02$
	eeog	eeog	eeog
	avg: 470 med: 57.5	avg: 1000 med: 1088	avg: 1000 med: 1142

Table 2: Duopoly Experiment Needle In Haystack

	k = 20	k = 250	k = 500
TS vs DG	<b>0.64</b> $\pm 0.03$	<b>0.6</b> $\pm 0.03$	<b>0.64</b> $\pm 0.03$
	eeog	eeog	eeog
	avg: 200 med: 27	avg: 370 med: 0	avg: 580 med: 121.5
TS vs DEG	<b>0.57</b> $\pm 0.03$	<b>0.52</b> $\pm 0.03$	<b>0.56</b> $\pm 0.02$
	eeog	eeog	eeog
	avg: 150 med: 14	avg: 460 med: 78.5	avg: 740 med: 627.5
DG vs DEG	<b>0.46</b> $\pm 0.03$	<b>0.42</b> $\pm 0.02$	<b>0.42</b> $\pm 0.02$
	eeog	eeog	eeog
	avg: 340 med: 128.5	avg: 650 med: 408	avg: 690 med: 466.5

The first line in each cell contains the average market share received by the firm playing Alg 1 (and the market share of Alg 2 is 1 - Alg 1 Market Share) as well as a 95 % confidence band. For example, the cell in the top left indicates that TS gets on average 64% of the market when played against DG. The next line contain the average and median effective end of game for this set of simulations.

**Temporary Monopoly** We now consider asymmetries in the timing of entry so that one firm enters the market before the other and serves as a monopolist in the periods until the other firm enters. In terms of our model, this corresponds to varying the value of  $X$ . For this section we report results fixing  $k = 20$  and focus on the priors where we observed that, for this warm start,  $DG > TS$ <sup>8</sup>. It seems clear that the temporary

<sup>7</sup>Interestingly, the EEOG also appears to be skewed to the right, similar to the distribution of reputation differences

<sup>8</sup>However, we report the results for the Needle In Haystack prior in the supplementary material. The results are unsurprising, where  $TS$  is the dominant strategy for both the entrant and incumbent.



monopoly will get a larger market share regardless of what algorithm it plays, but is the incumbent incentivized to commit to *TS*?

**Finding 5.** *Counterintuitively, allowing one firm to be a temporary monopolist for a sufficiently long time will induce that firm to commit to the better learning algorithm, *TS*, even if, under the permanent duopoly, they would commit to *DG**

In table 3 we can see that, for  $X = 200$ , *TS* is a dominant strategy for the incumbent and that *DG* is the dominant strategy for the entrant on the Heavy Tail prior <sup>9</sup>. In the supplementary material we report the same experiment for different values of  $X$  and show that our main result that, for sufficiently large  $X$ , *TS* is the dominant strategy for the incumbent is robust across priors.

Table 3: Temporary Monopoly Heavy Tail  $X = 200$

	TS	DEG	DG
	<b>0.003</b> $\pm 0.003$	<b>0.083</b> $\pm 0.02$	<b>0.17</b> $\pm 0.02$
TS	Var:0.002 ES:100%	Var:0.07 ES:97%	Var:0.1 ES:95%
	<b>0.045</b> $\pm 0.01$	<b>0.25</b> $\pm 0.02$	<b>0.23</b> $\pm 0.02$
DEG	Var:0.03 ES:92%	Var:0.1 ES:75%	Var:0.1 ES:78%
	<b>0.12</b> $\pm 0.02$	<b>0.36</b> $\pm 0.03$	<b>0.3</b> $\pm 0.02$
DG	Var:0.08 ES:88%	Var:0.2 ES:76%	Var:0.1 ES:64%

The columns represent the strategy of the incumbent and the rows represent the strategy of the entrant. The first line in each cell contains the average market share for the entrant over  $N = 1000$  simulations as well as a 95% confidence interval. The second line contains the sample variance of the observed market shares and the third line contains the fraction of simulations that ended up with one firm getting  $> 90\%$  of the market. Note that smaller values in the table are better for the incumbent. Market shares are calculated as the fraction of users selecting a particular firm *after* the entrant has already entered (i.e. the free rounds to firm 2 do not count towards the share)

Why do we observe that *TS* is the dominant strategy for the incumbent whereas in the permanent duopoly experiment we saw that, under the same other conditions, *DG* was preferred to *TS*? The intuition for this is that competition in the duopoly forces the firms to worry about their reputation which dissuades them from committing to algorithms that involve pure exploration in the early rounds. This intuition is very similar to that observed in the previous section by increasing warm start. One can view allowing one firm to temporarily be a monopolist as temporarily relaxing the "incentive" component of exploration, exploitation, and incentives so that the incumbent firm only faces the classic tradeoff between exploration and exploitation. The incumbent only needs to worry about her reputation after  $X$  periods when the entrant comes into the market and again

<sup>9</sup>The result for the Uniform is the same, though it takes a higher  $X$  to induce the incumbent to play *TS*. However, while  $DG > TS$  for the entrant, *DG* does roughly the same as *DEG* on the Uniform prior. The results are in the supplementary material.

needs to worry about incentivizing agents to select them over their competition. As a result, the incumbent is incentivized to commit to an algorithm that does exploration in the early rounds since she no longer suffers the same relative reputational cost that she would suffer under competition as long as the  $X$  is sufficiently large that she can begin to recover the reputational costs of exploration. Thus, counterintuitively, by having one firm be a monopoly and dominate the market, we can incentivize them to play *TS*.

## Data and Reputation as Barriers to Entry

An alternative interpretation of the results in the previous section is that the "temporary monopoly" provides a first mover advantage to the incumbent firm and that this first mover advantage allows the firm to both get a data advantage as well as a reputational advantage over the entrant. This provides an alternative interpretation of the classic "first mover advantage" that has been well-studied in economics and marketing (Kerin, Varadarajan, and Peterson, 1992) where, in our model, an incumbent can use data and reputation as a barrier to entry simply due to the fact that they were in the market before the entrant. Which plays a bigger role in preventing the entrant from being able to establish market share? We run two additional experiments, modifying the previous temporary monopoly experiment so that in one experiment the reputation of the incumbent is artificially erased (Reputation Erased Experiment) and another in which the information gained by the incumbent is artificially erased so that the posterior is reset to the prior (Information Erased Experiment).

Table 4: Reputation Erased Experiment, Heavy Tail

	TS	DEG	DG
	<b>0.016</b> $\pm 0.0075$	<b>0.13</b> $\pm 0.02$	<b>0.2</b> $\pm 0.024$
TS	Var: 0.01 ES: 100 %	Var: 0.1 ES: 97 %	Var: 0.1 ES: 96 %
	<b>0.068</b> $\pm 0.013$	<b>0.29</b> $\pm 0.024$	<b>0.26</b> $\pm 0.024$
DEG	Var: 0.05 ES: 93 %	Var: 0.2 ES: 75 %	Var: 0.2 ES: 80 %
	<b>0.15</b> $\pm 0.019$	<b>0.38</b> $\pm 0.028$	<b>0.33</b> $\pm 0.024$
DG	Var: 0.1 ES: 87 %	Var: 0.2 ES: 80 %	Var: 0.2 ES: 67 %

**Finding 6.** *When the incumbent plays *TS*, the data advantage serves as a larger barrier to entry compared to reputation. However, when the incumbent plays *DEG* or *DG* there is no substantial difference between the two.*

Tables 4 and 5 display the results of these experiments. Maintaining either the data or reputation advantage alone still allows the incumbent to retain a significant portion of the market. The main interesting finding here, which is robust across priors for this parameterization, is that information serves as a larger barrier to entry than

Table 5: Information Erased Experiment, Heavy Tail

	TS	DEG	DG
	<b>0.024</b> $\pm$ 0.0094	<b>0.16</b> $\pm$ 0.022	<b>0.22</b> $\pm$ 0.025
TS	Var: 0.02	Var: 0.1	Var: 0.2
	ES: 100 %	ES: 97 %	ES: 95 %
	<b>0.24</b> $\pm$ 0.025	<b>0.29</b> $\pm$ 0.024	<b>0.27</b> $\pm$ 0.024
DEG	Var: 0.2	Var: 0.1	Var: 0.1
	ES: 94 %	ES: 72 %	ES: 76 %
	<b>0.33</b> $\pm$ 0.028	<b>0.38</b> $\pm$ 0.026	<b>0.33</b> $\pm$ 0.023
DG	Var: 0.2	Var: 0.2	Var: 0.1
	ES: 94 %	ES: 74 %	ES: 58 %

reputation when the incumbent plays *TS*. One possible intuition for this is that, for Heavy Tail, the information acquired after the  $X$  rounds under *TS* is sufficient so that it becomes relatively easier for *TS* to recover the reputation loss compared to the information loss, but since *DEG* and *DG* learn slower and they’ve acquired less information over the  $X$  rounds relative to *TS* there is less of an informational advantage.

Though the setup is purely experimental, it is nonetheless interesting to look at if the same strategies remain best responses in this setting compared to the setting where both data and reputation are retained. We find that the best responses remain the same for the Heavy Tail prior, but for the Uniform prior they change where *TS* is weakly dominant for the incumbent under the reputation erased treatment but not under the information erased treatment.

## References

- Aghion, P.; Bloom, N.; Blundell, R.; Griffith, R.; and Howitt, P. 2005. Competition and innovation: An inverted-u relationship. *The Quarterly Journal of Economics* 120(2):701–728.
- Agrawal, S., and Goyal, N. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, 39–1.
- Barro, R. J., and Sala-i Martin, X. 2004. Economic growth: Mit press. *Cambridge, Massachusetts*.
- Bolton, P., and Harris, C. 1999. Strategic experimentation. *Econometrica* 67(2):349–374.
- Bubeck, S.; Cesa-Bianchi, N.; et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5(1):1–122.
- Che, Y.-K., and Hörner, J. 2017. Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics* 133(2):871–925.
- Immorlica, N.; Kalai, A. T.; Lucier, B.; Moitra, A.; Postlewaite, A.; and Tennenholtz, M. 2011. Dueling algorithms. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, 215–224. ACM.
- Keller, G.; Rady, S.; and Cripps, M. 2005. Strategic experimentation with exponential bandits. *Econometrica* 73(1):39–68.
- Kerin, R. A.; Varadarajan, P. R.; and Peterson, R. A. 1992. First-mover advantage: A synthesis, conceptual framework, and research propositions. *The Journal of Marketing* 33–52.
- Kremer, I.; Mansour, Y.; and Perry, M. 2014. Implementing the “wisdom of the crowd”. *Journal of Political Economy* 122(5):988–1012.
- Mansour, Y.; Slivkins, A.; and Syrgkanis, V. 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 565–582. ACM.
- Mansour, Y.; Slivkins, A.; and Wu, Z. S. 2018. Competing bandits: Learning under competition. *Innovations in Theoretical Computer Science (ITCS)*.
- Schumpeter, J. A. 2010. *Capitalism, socialism and democracy*. Routledge.
- Slivkins, A. 2017. An introduction to multi-armed bandits.
- Tirole, J. 1988. *The theory of industrial organization*. MIT press.