

# Competing Bandits

June 25, 2018

## Overview of Old Simulation Results

First, summarizing the simulation results that we have seen:

### Priors:

1. Needle In a Haystack ( $x, K$ ) = ( $K-1$ ) arms with a mean of  $x$  and 1 arm with mean  $x + 0.2$
2. Heavy Tail - means of arms randomly drawn from  $\text{beta}(0.6, 0.6)$
3. Uniform - means of arms randomly drawn from  $[0.25, 0.75]$
4. .5/.7 Prior - means of arms randomly drawn from 0.5, 0.7

**Question:** What algorithms win when both firms start at the same time (simultaneous entry)?  
How does this vary across agent algorithms?

### 1. For HM

- (a)  $TS > DEG$ ,  $DG$  for Needle In Haystack (0.5, 10), (0.7, 10)
- (b) 50/50 for Needle In Haystack (0.5, 3), (0.1, 10), (0.3, 10), Uniform ( $K = 3, 10$  - low warm start), .5/.7 prior
- (c)  $DEG, DG > TS$  for Heavy Tail ( $K = 3, 10$ ), Uniform ( $K = 10$  - moderate warm start of 15, 20)
- (d)  $DG > DEG$  for Heavy Tail (this holds for really large time horizons!)
- (e) Results sensitive to the warm start (see latest experiments)
- (f) Effective End of Game is skewed to the right with many instances having a median of 0, which means that many of the games end simply from the choices made in the warm start, but a relatively larger mean.
- (g) Many extreme shares - one principal takes most of the market

### 2. For HMR:

- (a)  $TS > DEG$ ,  $DG$  wins with sufficiently large time horizon
- (b) One exception is the .5/.7 prior (where means of the arms are randomly drawn from 0.5, 0.7) where even this is 50/50
- (c) We see that  $DG > DEG$  for HeavyTail prior, with  $K = 3$

(d) Moderate variance in the market shares

3. For SM

(a) Qualitatively similar to HMR where  $TS > DEG, DG$  except in the .5/.7 prior.

(b) However,  $DG > DEG$  does not hold for HeavyTail prior.

(c) Very low variance in the shares

**Question:** What happens if there is asymmetric entry so that one principal has an incumbency advantage? When the other principal enters the incumbent has both an informational and reputation advantage. Which serves as bigger barriers to entry?

1. For the incumbent TS is a dominant strategy across priors and agent models that we have tested out.
2. For SM, TS is not better by much
3. For HMR, TS is better by more than SM
4. For HM, TS is better by more than HMR
5. For the entrant, it is ambiguous still what is the best strategy

**Information or reputation?**

1. For HM erasing reputation hurts incumbent more (reputational advantage is more important than data advantage)
2. For HMR/SM - erasing either hurts a bit, but erasing both hurts a lot implying that reputational and data advantage substitute for one another

## Conjectures

Our goal in this section is to be able to have a coherent story to explain the results discussed in the previous section as well as guide what we ought to validate in order to confirm the story. In our model we introduce a notion of reputation in the competing bandits game by allowing agents to learn about the past performance of the firms. The consequence of this for the firms is that now the principals face a dilemma between exploration and reputation since exploration today may give me better information for tomorrow but may also hurt my reputation for tomorrow so that I lose users in competition tomorrow while exploration may lead me to not gain much information for tomorrow but may help my reputation for tomorrow. This tradeoff is especially pertinent when just starting out and neither principal has strong reputation nor sufficient information so that there is a value to exploration.

This leads to following question: Under what conditions are principals incentivized to adopt smarter learning algorithms?

To begin, focus on a standard multi-armed bandit problem. If we compare the performance of a greedy algorithm and an adaptive exploration algorithm in isolation, then on sufficiently hard learning problems where the adaptive exploration algorithm has to explore for a while, we expect the performance of the greedy algorithm to be better than the adaptive exploration algorithm in the early rounds and the performance of the adaptive exploration algorithm to be better than the greedy algorithm in the later rounds. Thus, for a given learning problem, there is some threshold  $\bar{t}$  where, if the instance runs for  $T > \bar{t}$  then the adaptive exploration algorithm would, on average, perform better. Additionally, we expect that the “performance gap” between a “good” algorithm such as TS and a “bad” algorithm such as DG will be greater for “harder” learning instances.

When we move to the competition game in our model we introduce a notion of reputation so that suboptimal exploration does not only impact my reward today but my reputation tomorrow. In comparing the performance of the algorithms in isolation we only had what was effectively a “learning” phase and an “exploitation” phase. When adding reputation we conjecture that we can view this as adding a third phase in between learning and exploitation which is a reputation-recovery phase. The main effect that this has is that the threshold at which a “better” algorithm is expected to “win” in competition with reputation is greater than  $\bar{t}$  from before since it is not only necessary to have better instantaneous regret than the other algorithm but the algorithm also needs additional rounds to recover the reputation loss suffered from exploration (let’s call this time  $t^w(B)$ ).

**Main Conjecture:** Fix a bandit instance  $B$ , then for that bandit instance there is some  $t^w(B)$  where a “better” algorithm will on average win so long as the time horizon is sufficiently long and somehow it eventually gets  $t^w(B)$  observations.

In this context, we for now say an algorithm  $a$  is “better” than an algorithm  $b$  on  $B$  if, in isolation, past some  $t^w$ , algorithm  $a$  gets a weakly higher reputation  $b$  more than 50% of the time when run on  $I^1$ .

Further, we conjecture that  $t^w(B)$  is increasing in the difficulty of the learning problem. We conjecture that the hardness,  $H$ , of an instance can be defined by

$$H = \mathbb{E} \left[ \sum_{\mu_i < \mu^*} \frac{1}{\mu^* - \mu_i} \right], \text{ where } \mu^* \text{ is the mean of the best arm.}$$

---

<sup>1</sup>Note that previously we were looking at the mean in the preliminary plots, see later discussion. It would be better to have a definition that is not dependent on relative performance but just on the instance itself

There are several ways in our model in which a principal playing a good algorithm can get  $t^w(B)$  observations:

1. There is any randomness in the agent response function, such as the HardMaxWithRandom or SoftMax agent response function. Depending on the parameterization, at some point the good algorithm should get enough random agents to pass the threshold and “win” the game on average.
2. A sufficiently long incumbency advantage (not sure what sufficiently long is here)
3. A sufficiently long “warm start” where principals get free users (“sufficiently long” here should be  $t^w(B)$  especially since the median EEOG is 0)

Thus, whether better algorithms do better in competition or not depends on whether there is any randomness in the agent response rule or how many free observations we give the firms at the start of the game. Counterintuitively, under the HardMax decision rule, when firms enter at the same time they are *not* incentivized to play the better algorithm when there are reputational costs to exploration. However, if we give the firms a sufficient number of rounds as a monopolist or give them a sufficient number of free customers then they *are* incentivized to play the better algorithm. It’s interesting to relate this to the R & D literature where generally the development of R & D is costly but the ability to patent after invention allows the firms to extract monopoly profits ex-post and thus recoup the R & D costs and this incentives R & D ex-ante. In our model we have no explicit development costs but rather have *reputational* costs due to the exploration involved in better algorithms. Instead of allowing the firm to recoup the R & D costs *after* invention as a monopolist via patents we have that it is better for the firm to be a monopolist *at the start* sufficiently long to recoup the costs of exploration while having the threat that there will be an entrant at some fixed point in the future. In some sense we can view the “monopolist” or “free consumer” case here as saying that we relax the incentive constraint on the consumer in the early rounds in order to incentivize the firm to employ a better algorithm when the incentive constraints for the consumer kick back in when the entrant comes into the market.

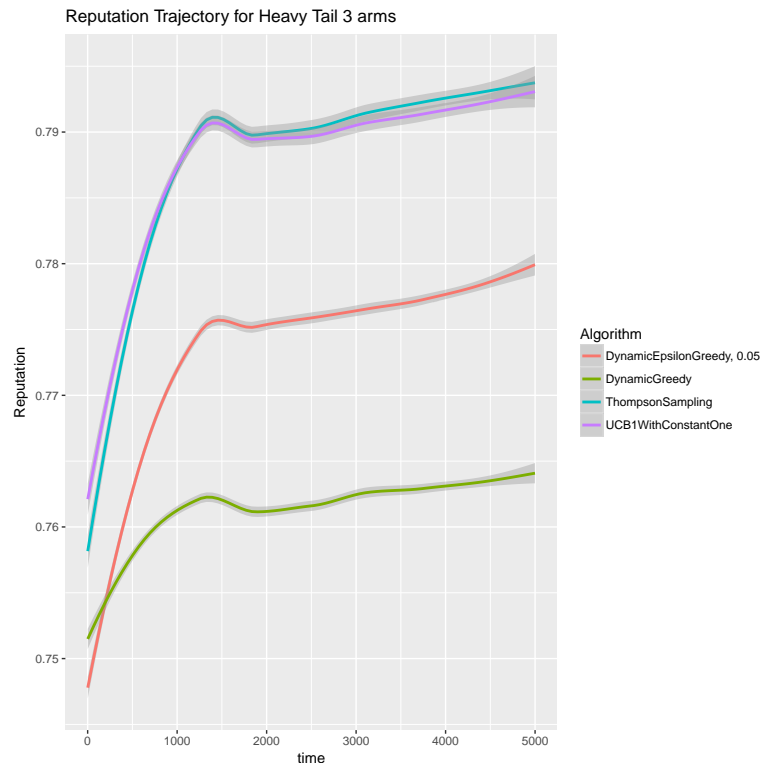
Crucially, this result comes from the fact that firms cannot observe the behavior of other firms and learn from the behavior of other firms since there are two advantages that the initial “free” rounds gives the firms: reputation and information. At the start of the game there is a tradeoff between exploration and reputation. Initial exploration leads to bad initial reputation but improved information. Since consumers are myopic and don’t care about the information gain of the firm but only the reputation of the firm, the firm employing algorithms that involve purposeful exploration suffer but with sufficiently many initial free rounds, the firm will be able to recoup these reputation losses and thus will have both better information and reputation than an entrant. If the entrant could observe the actions of the other firm, then the entrant could potentially use this as information about the bandit instance and learn not only from consumers but from the other firms. However, we require that firms can only directly learn from consumers that pick them. Thus, upon entry, the information gained by the incumbent is the incumbent’s alone and thus the incumbency gives the incumbent a data advantage.

A natural follow-up question is then, conditional on having a firm early in the market, what serves as stronger barriers to entry in our model: reputation or data advantage? Our results show that, for HardMax, the reputation advantage is more useful but that the data advantage still is significant relative to the results from the original game. I don’t really have a good intuition for

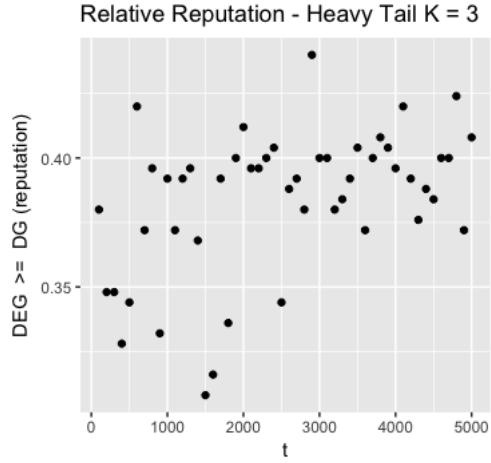
why this should be the case, but wonder if this has anything to do with good information can't help you if you get unlucky with your rewards.

## DEG vs DG

The first thing that is at odds with the above conjectures is that we've observed that  $DEG > DG$  for Heavy Tail. However, the preliminary plots that we had been looking at were looked at the reputation curves for the *mean* reputation across simulations. For instance, the mean plot for Heavy Tail looked as follows:

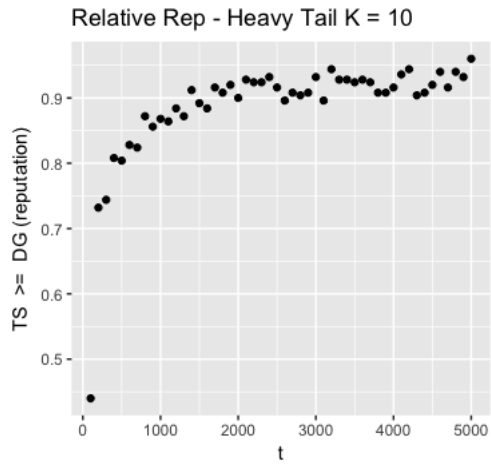


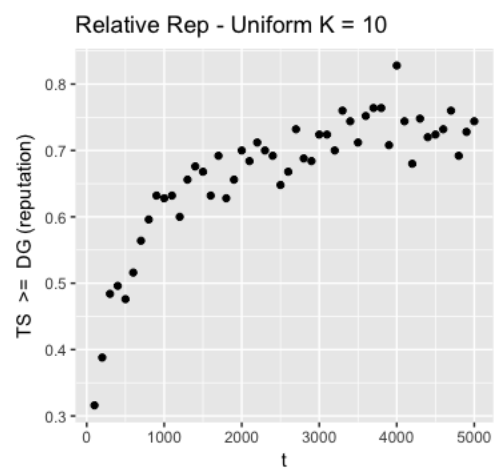
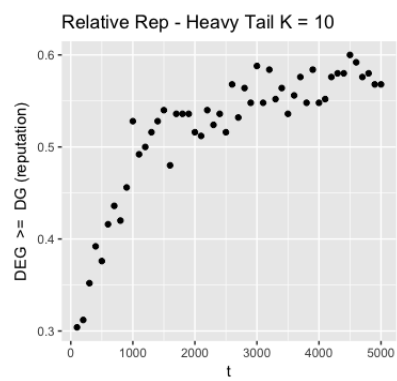
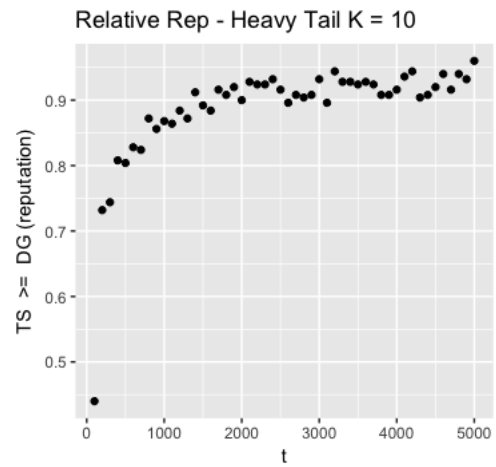
but if, instead of looking at mean reputation, we look at the fraction of rounds where the reputation of DEG is higher than DG we see that  $DG > DEG$  on this instance.

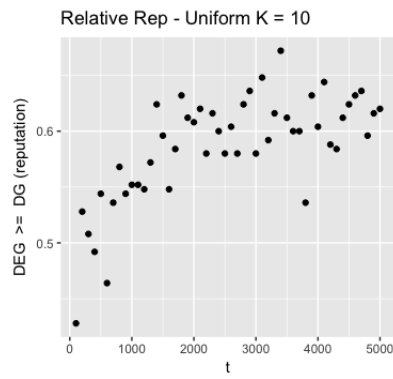
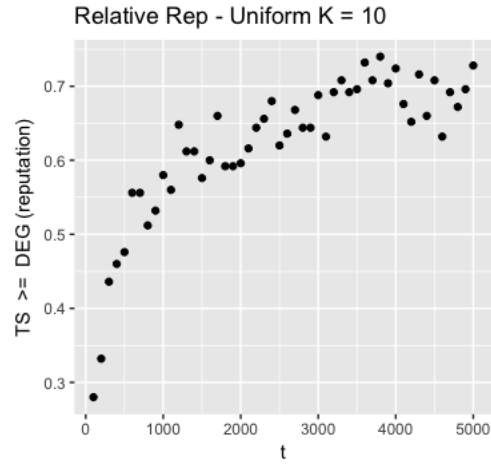


## Large Warm Start

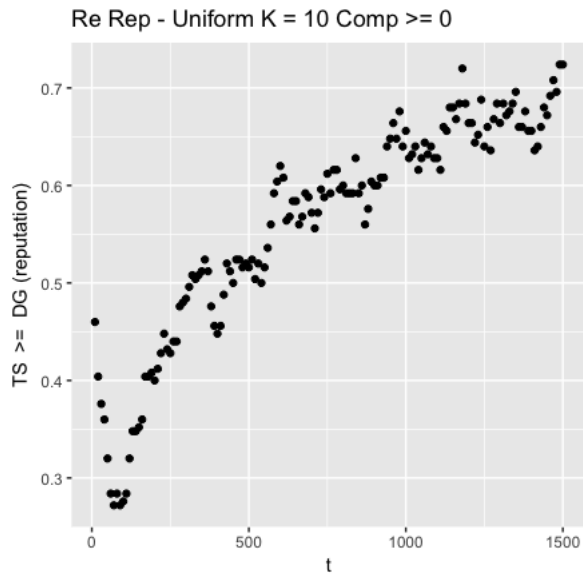
We want to see if, for sufficiently large warm start, we indeed see that TS wins. First, let's take a look at the relative reputation plots for Heavy Tail and Uniform. Notice that, for both,  $DEG, DG > TS$  in the earlier time horizons and  $TS > DEG, DG$  in the later time horizons, with Uniform taking longer for  $TS > DEG, DG$  than Heavy Tail.



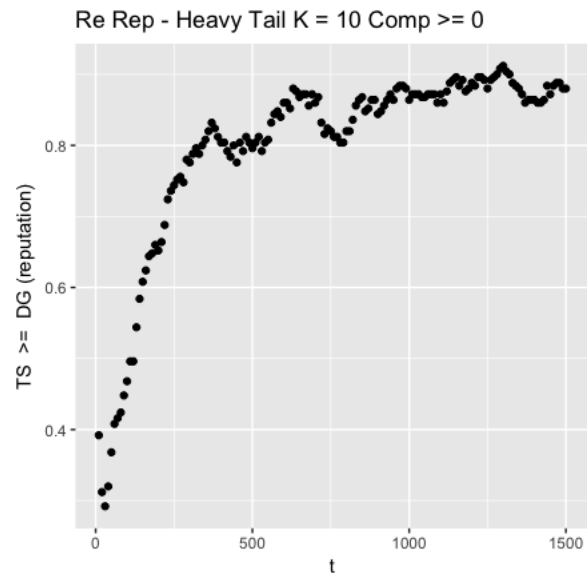




Given this we conjectured that in the competition game,  $DEG, DG > TS$  in the early rounds but  $TS > DEG, DG$ . This largely seems to be the case. See the results below. Additionally, here are two more preliminary plots with smaller time gaps.







## Results for Heavy Tail HardMax K=10

	WS = 5	WS = 20	WS = 50	WS = 100	WS = 200	WS = 400	WS = 700	WS = 1000
TS vs DG	<b>0.37</b> (0.05) <u>eeog</u> avg: 31 med: 0	<b>0.32</b> (0.05) <u>eeog</u> avg: 150 med: 0	<b>0.39</b> (0.05) <u>eeog</u> avg: 220 med: 0	<b>0.55</b> (0.05) <u>eeog</u> avg: 300 med: 0	<b>0.73</b> (0.04) <u>eeog</u> avg: 1400 med: 226.5	<b>0.77</b> (0.04) <u>eeog</u> avg: 1400 med: 375	<b>0.77</b> (0.04) <u>eeog</u> avg: 1700 med: 705	<b>0.81</b> (0.03) <u>eeog</u> avg: 1400 med: 0
TS vs DEG	<b>0.36</b> (0.05) <u>eeog</u> avg: 10 med: 0	<b>0.34</b> (0.05) <u>eeog</u> avg: 36 med: 0	<b>0.43</b> (0.06) <u>eeog</u> avg: 160 med: 0	<b>0.65</b> (0.05) <u>eeog</u> avg: 330 med: 0	<b>0.91</b> (0.02) <u>eeog</u> avg: 1100 med: 107.5	<b>0.93</b> (0.02) <u>eeog</u> avg: 1200 med: 183	<b>0.9</b> (0.02) <u>eeog</u> avg: 1500 med: 1235.5	<b>0.92</b> (0.02) <u>eeog</u> avg: 1400 med: 878
DG vs DEG	<b>0.59</b> (0.05) <u>eeog</u> avg: 470 med: 2	<b>0.62</b> (0.05) <u>eeog</u> avg: 800 med: 10	<b>0.62</b> (0.05) <u>eeog</u> avg: 1300 med: 397.5	<b>0.6</b> (0.05) <u>eeog</u> avg: 1500 med: 866.5	<b>0.62</b> (0.05) <u>eeog</u> avg: 1900 med: 1627.5	<b>0.55</b> (0.05) <u>eeog</u> avg: 1700 med: 1338	<b>0.54</b> (0.05) <u>eeog</u> avg: 1600 med: 1056.5	<b>0.53</b> (0.05) <u>eeog</u> avg: 1500 med: 972.5

[1] “

”

## Results for Uniform HardMax K=10

	<b>WS = 5</b>	<b>WS = 20</b>	<b>WS = 50</b>	<b>WS = 100</b>	<b>WS = 200</b>	<b>WS = 400</b>	<b>WS = 700</b>	<b>WS = 1000</b>
TS vs DG	<b>0.45</b> (0.06) <u>eeog</u> avg: 220 med: 0	<b>0.46</b> (0.05) <u>eeog</u> avg: 500 med: 1	<b>0.48</b> (0.05) <u>eeog</u> avg: 1000 med: 55.5	<b>0.45</b> (0.05) <u>eeog</u> avg: 1300 med: 528	<b>0.53</b> (0.05) <u>eeog</u> avg: 1900 med: 1596.5	<b>0.61</b> (0.04) <u>eeog</u> avg: 2100 med: 2182	<b>0.65</b> (0.04) <u>eeog</u> avg: 2300 med: 2422.5	<b>0.67</b> (0.04) <u>eeog</u> avg: 2200 med: 2182.5
TS vs DEG	<b>0.49</b> (0.06) <u>eeog</u> avg: 340 med: 0	<b>0.39</b> (0.05) <u>eeog</u> avg: 490 med: 7	<b>0.38</b> (0.05) <u>eeog</u> avg: 1000 med: 127.5	<b>0.43</b> (0.05) <u>eeog</u> avg: 1500 med: 837.5	<b>0.52</b> (0.05) <u>eeog</u> avg: 1900 med: 1804	<b>0.56</b> (0.04) <u>eeog</u> avg: 2400 med: 2346.5	<b>0.59</b> (0.04) <u>eeog</u> avg: 2600 med: 2804.5	<b>0.57</b> (0.04) <u>eeog</u> avg: 2600 med: 2805
DG vs DEG	<b>0.46</b> (0.05) <u>eeog</u> avg: 690 med: 8	<b>0.49</b> (0.05) <u>eeog</u> avg: 1200 med: 280.5	<b>0.49</b> (0.04) <u>eeog</u> avg: 1900 med: 1473	<b>0.49</b> (0.04) <u>eeog</u> avg: 2300 med: 2623	<b>0.48</b> (0.04) <u>eeog</u> avg: 2400 med: 2687.5	<b>0.43</b> (0.04) <u>eeog</u> avg: 2600 med: 2743.5	<b>0.38</b> (0.04) <u>eeog</u> avg: 2600 med: 2879	<b>0.4</b> (0.04) <u>eeog</u> avg: 2500 med: 2726

[1] “

”

## Learning Complexity

Using the same instances that we used for the plots generated in the preliminary plots, we calculate the empirical “learning complexity” for the  $K = 10$  instances according to the hardness metric defined above.

Needle In Haystack - 0.5

**Mean:** 45 **Median :** 45

Heavy Tail

**Mean:** 196.897 **Median :** 49.08325

.5/.7 Random Draw

**Mean:** 23.82 **Median:** 25

Uniform

**Mean:** 171.7698 **Median :** 79.59571

This is promising given that Heavy Tail and Uniform are the two instances where we see that  $DEG, DG > TS$  for HardMax since before we conjectured that on harder learning instances, for small enough warm start observations, TS will do purposeful exploration and get bad reputation and thus lose the competition game. There is probably more we should validate about this metric but need to think about it more.

One puzzling thing about it is, for instance, on the easy instance of 0.5 / 0.7 Random Draw, we see that things are roughly 50 / 50 but we see that  $TS > DEG, DG$  on Needle In Haystack - 0.5? One idea is that on 0.5 / 0.7 learning is so easy that there is no value to smarter exploration and that on Needle In Haystack learning is not trivial but is very quick so that only a few samples are needed and thus the reputational cost is negligible but since  $DG$  can’t trivially find the best arm there is value to smarter exploration. Not sure though.

## Incumbent Experiment - Varying the number of free observations

Another implication of the proposed conjectures is that TS ought to be the dominant strategy for the incumbent as long as they are an incumbent for sufficiently long before the entrant enters the market. We fix the agent response model as HardMax. Then, for sufficiently hard learning problems it ought to be the case that if the incumbent only has relatively few periods as the monopolist then TS will not necessarily win. However, with sufficiently long time as the monopolist then TS ought to win.

We verify this in two experiments by first looking where the incumbent gets 50 free observations and then where the incumbent gets 200 free observations. We observe that in the case of 50 free observations, TS does not beat DEG and DG from the “harder” instances defined by the complexity metric defined above (i.e. Uniform and Heavy Tail), wins in the relatively easier but not trivial learning problem NIH, and algorithm choice makes little difference in the 0.5 / 0.7 Prior.

When we increase the number of free observations to 200 we see that TS is a better algorithm choice for the incumbent (one exception - Uniform against DEG but this is likely due to the fact that we need more free observations, see the preliminary plots).

Results for HardMax  $t = 5000$  Needle In Haystack - 0.5 50 free obs  $K = 10$ 

	Thompson Sampling	Dynamic Epsilon Greedy	DynamicGreedy
Thompson Sampling	<b>0.18</b> +/- 0.042 0.14 (0.12, 0.16) Extreme Shares: 99 %	<b>0.27</b> +/- 0.049 0.19 (0.16, 0.22) Extreme Shares: 98 %	<b>0.37</b> +/- 0.052 0.21 (0.18, 0.25) Extreme Shares: 93 %
Dynamic Epsilon Greedy	<b>0.1</b> +/- 0.033 0.082 (0.07, 0.097) Extreme Shares: 98 %	<b>0.19</b> +/- 0.042 0.14 (0.12, 0.16) Extreme Shares: 93 %	<b>0.33</b> +/- 0.05 0.19 (0.16, 0.23) Extreme Shares: 87 %
DynamicGreedy	<b>0.054</b> +/- 0.024 0.043 (0.037, 0.051) Extreme Shares: 98 %	<b>0.11</b> +/- 0.031 0.076 (0.065, 0.09) Extreme Shares: 90 %	<b>0.25</b> +/- 0.042 0.14 (0.12, 0.16) Extreme Shares: 79 %

Results for HardMax  $t = 5000$  Heavy Tail 50 free obs  $K = 10$ 

	Thompson Sampling	Dynamic Epsilon Greedy	DynamicGreedy
Thompson Sampling	<b>0.023</b> +/- 0.017 0.023 (0.02, 0.027) Extreme Shares: 100 %	<b>0.046</b> +/- 0.024 0.043 (0.037, 0.051) Extreme Shares: 100 %	<b>0.11</b> +/- 0.034 0.091 (0.078, 0.11) Extreme Shares: 97 %
Dynamic Epsilon Greedy	<b>0.17</b> +/- 0.042 0.14 (0.12, 0.16) Extreme Shares: 99 %	<b>0.13</b> +/- 0.034 0.088 (0.075, 0.1) Extreme Shares: 88 %	<b>0.14</b> +/- 0.036 0.098 (0.084, 0.12) Extreme Shares: 90 %
DynamicGreedy	<b>0.22</b> +/- 0.046 0.17 (0.14, 0.2) Extreme Shares: 99 %	<b>0.2</b> +/- 0.042 0.14 (0.12, 0.16) Extreme Shares: 91 %	<b>0.16</b> +/- 0.033 0.082 (0.07, 0.097) Extreme Shares: 74 %

Results for HardMax  $t = 5000$  .5/.7 Random Draw 50 free obs  $K = 10$ 

	Thompson Sampling	Dynamic Epsilon Greedy	DynamicGreedy
Thompson Sampling	<b>0.19</b> +/- 0.038 0.11 (0.097, 0.13) Extreme Shares: 82 %	<b>0.17</b> +/- 0.037 0.1 (0.09, 0.12) Extreme Shares: 82 %	<b>0.21</b> +/- 0.04 0.12 (0.11, 0.15) Extreme Shares: 82 %
Dynamic Epsilon Greedy	<b>0.25</b> +/- 0.042 0.14 (0.12, 0.16) Extreme Shares: 77 %	<b>0.24</b> +/- 0.04 0.12 (0.11, 0.15) Extreme Shares: 73 %	<b>0.22</b> +/- 0.039 0.12 (0.099, 0.14) Extreme Shares: 72 %
DynamicGreedy	<b>0.28</b> +/- 0.045 0.16 (0.13, 0.19) Extreme Shares: 80 %	<b>0.2</b> +/- 0.036 0.098 (0.084, 0.12) Extreme Shares: 71 %	<b>0.22</b> +/- 0.038 0.11 (0.097, 0.13) Extreme Shares: 73 %

Results for HardMax t = 5000 Uniform 50 free obs K = 10

	Thompson Sampling	Dynamic Epsilon Greedy	DynamicGreedy
Thompson Sampling	<b>0.13</b> +/- 0.037 0.1 (0.089, 0.12) Extreme Shares: 98 %	<b>0.16</b> +/- 0.039 0.12 (0.1, 0.14) Extreme Shares: 95 %	<b>0.15</b> +/- 0.037 0.11 (0.093, 0.13) Extreme Shares: 92 %
Dynamic Epsilon Greedy	<b>0.25</b> +/- 0.047 0.17 (0.15, 0.2) Extreme Shares: 93 %	<b>0.19</b> +/- 0.04 0.12 (0.11, 0.14) Extreme Shares: 85 %	<b>0.21</b> +/- 0.04 0.12 (0.1, 0.14) Extreme Shares: 80 %
DynamicGreedy	<b>0.24</b> +/- 0.045 0.16 (0.14, 0.19) Extreme Shares: 90 %	<b>0.16</b> +/- 0.036 0.099 (0.085, 0.12) Extreme Shares: 84 %	<b>0.2</b> +/- 0.037 0.1 (0.089, 0.12) Extreme Shares: 73 %

Results for HardMax t = 5000 Needle In Haystack - 0.5 200 free obs K = 10

	Thompson Sampling	Dynamic Epsilon Greedy	DynamicGreedy
Thompson Sampling	<b>0.083</b> +/- 0.03 0.069 (0.06, 0.082) Extreme Shares: 99 %	<b>0.19</b> +/- 0.044 0.15 (0.13, 0.17) Extreme Shares: 98 %	<b>0.34</b> +/- 0.051 0.2 (0.17, 0.24) Extreme Shares: 90 %
Dynamic Epsilon Greedy	<b>0.071</b> +/- 0.028 0.06 (0.051, 0.071) Extreme Shares: 98 %	<b>0.2</b> +/- 0.043 0.14 (0.12, 0.17) Extreme Shares: 93 %	<b>0.28</b> +/- 0.048 0.18 (0.15, 0.21) Extreme Shares: 87 %
DynamicGreedy	<b>0.028</b> +/- 0.016 0.021 (0.018, 0.025) Extreme Shares: 99 %	<b>0.1</b> +/- 0.032 0.077 (0.066, 0.091) Extreme Shares: 95 %	<b>0.22</b> +/- 0.041 0.13 (0.11, 0.15) Extreme Shares: 81 %

Results for HardMax t = 5000 Heavy Tail 200 free obs K = 10

	Thompson Sampling	Dynamic Epsilon Greedy	DynamicGreedy
Thompson Sampling	<b>0.0067</b> +/- 0.0092 0.0066 (0.0056, 0.0078) Extreme Shares: 100 %	<b>0.023</b> +/- 0.017 0.021 (0.018, 0.025) Extreme Shares: 99 %	<b>0.064</b> +/- 0.027 0.055 (0.047, 0.065) Extreme Shares: 97 %
Dynamic Epsilon Greedy	<b>0.024</b> +/- 0.015 0.017 (0.015, 0.02) Extreme Shares: 98 %	<b>0.13</b> +/- 0.034 0.088 (0.076, 0.1) Extreme Shares: 86 %	<b>0.14</b> +/- 0.036 0.099 (0.085, 0.12) Extreme Shares: 89 %
DynamicGreedy	<b>0.063</b> +/- 0.024 0.043 (0.037, 0.051) Extreme Shares: 93 %	<b>0.19</b> +/- 0.041 0.13 (0.11, 0.16) Extreme Shares: 91 %	<b>0.15</b> +/- 0.032 0.08 (0.069, 0.095) Extreme Shares: 77 %

Results for HardMax t = 5000 .5/.7 Random Draw 200 free obs K = 10

	Thompson Sampling	Dynamic Epsilon Greedy	DynamicGreedy
Thompson Sampling	<b>0.15</b> +/- 0.034 0.087 (0.075, 0.1) Extreme Shares: 81 %	<b>0.14</b> +/- 0.032 0.081 (0.069, 0.096) Extreme Shares: 81 %	<b>0.18</b> +/- 0.036 0.099 (0.085, 0.12) Extreme Shares: 76 %
Dynamic Epsilon Greedy	<b>0.2</b> +/- 0.038 0.11 (0.095, 0.13) Extreme Shares: 75 %	<b>0.15</b> +/- 0.032 0.081 (0.07, 0.096) Extreme Shares: 78 %	<b>0.22</b> +/- 0.037 0.11 (0.093, 0.13) Extreme Shares: 70 %
DynamicGreedy	<b>0.24</b> +/- 0.04 0.12 (0.11, 0.15) Extreme Shares: 74 %	<b>0.19</b> +/- 0.035 0.096 (0.083, 0.11) Extreme Shares: 74 %	<b>0.2</b> +/- 0.037 0.11 (0.092, 0.13) Extreme Shares: 74 %

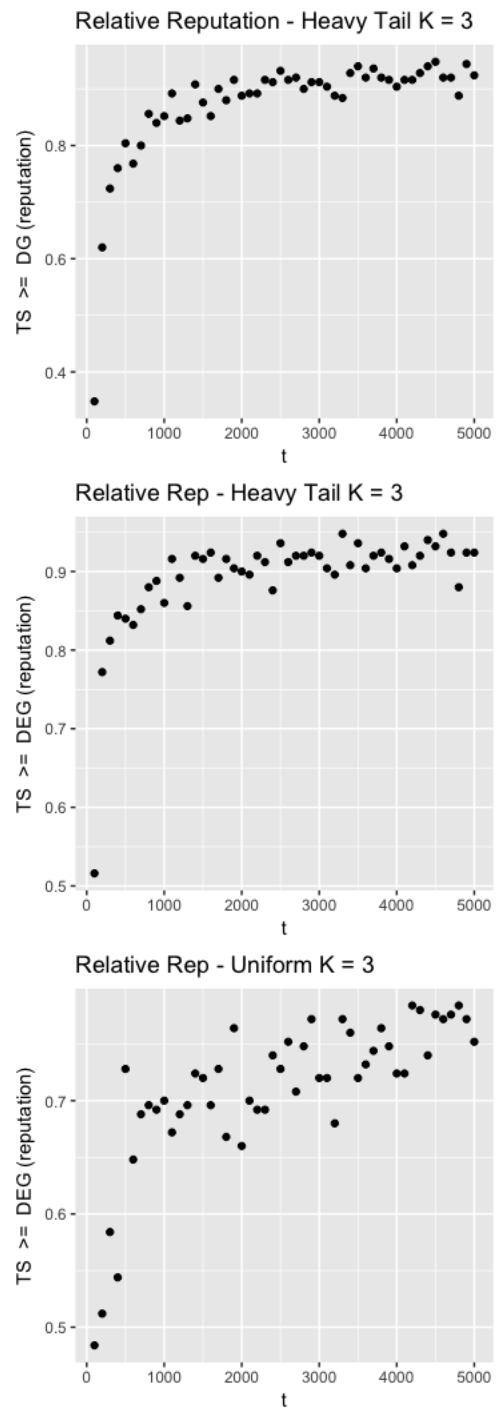
Results for HardMax t = 5000 Uniform 200 free obs K = 10

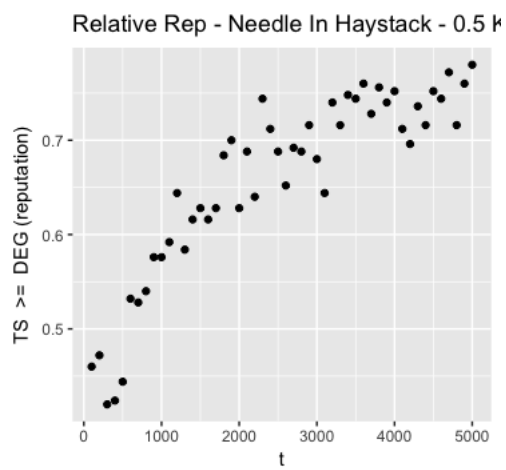
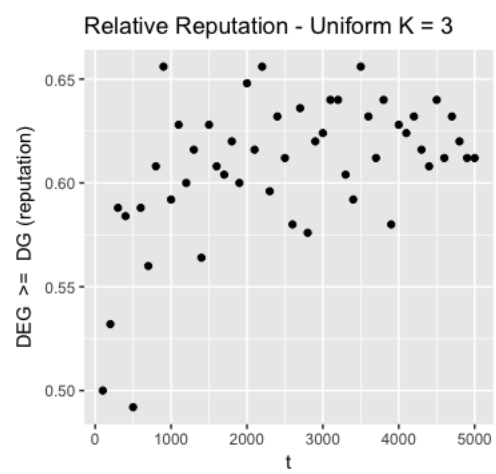
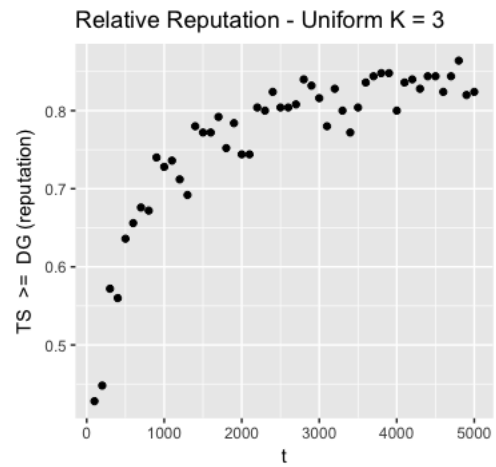
	Thompson Sampling	Dynamic Epsilon Greedy	DynamicGreedy
Thompson Sampling	<b>0.055</b> +/- 0.023 0.042 (0.036, 0.05) Extreme Shares: 95 %	<b>0.11</b> +/- 0.033 0.083 (0.071, 0.098) Extreme Shares: 93 %	<b>0.17</b> +/- 0.038 0.11 (0.098, 0.14) Extreme Shares: 88 %
Dynamic Epsilon Greedy	<b>0.13</b> +/- 0.035 0.093 (0.08, 0.11) Extreme Shares: 90 %	<b>0.12</b> +/- 0.031 0.077 (0.066, 0.091) Extreme Shares: 86 %	<b>0.19</b> +/- 0.039 0.12 (0.1, 0.14) Extreme Shares: 85 %
DynamicGreedy	<b>0.087</b> +/- 0.028 0.059 (0.051, 0.07) Extreme Shares: 92 %	<b>0.12</b> +/- 0.031 0.073 (0.063, 0.087) Extreme Shares: 85 %	<b>0.18</b> +/- 0.036 0.098 (0.084, 0.12) Extreme Shares: 77 %

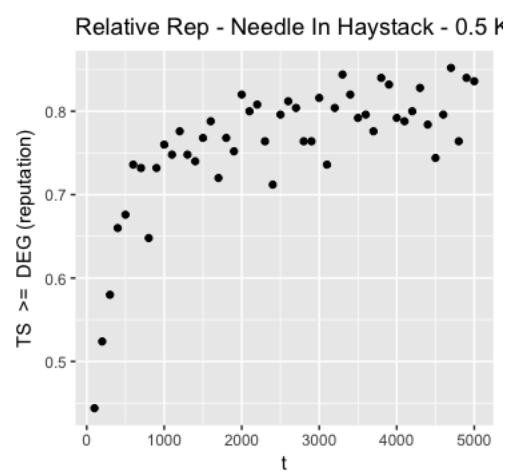
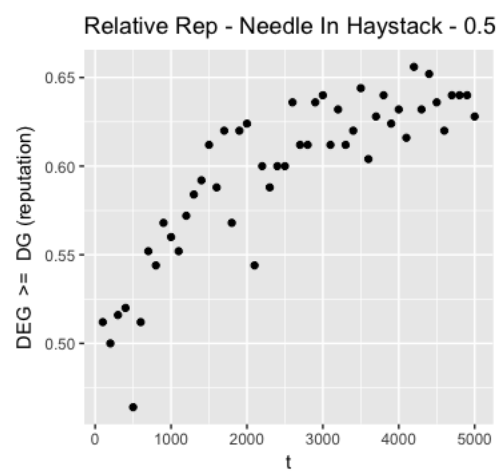
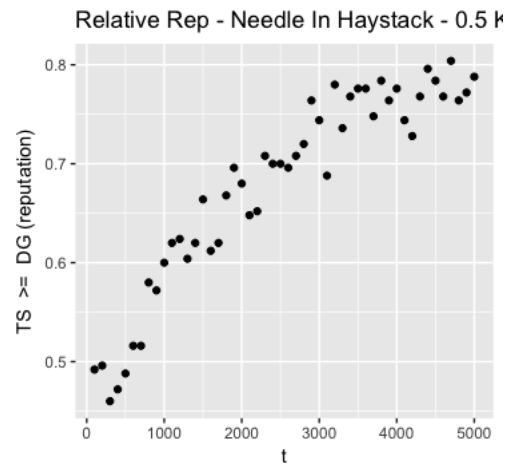


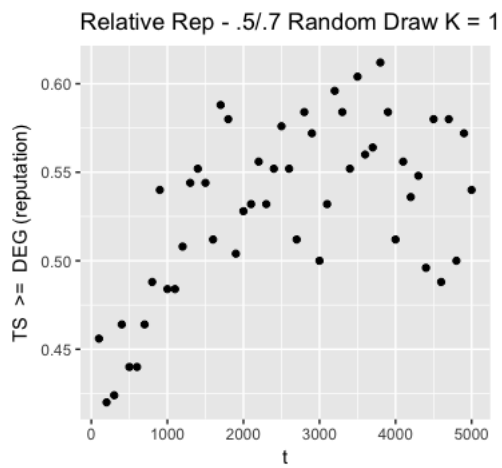
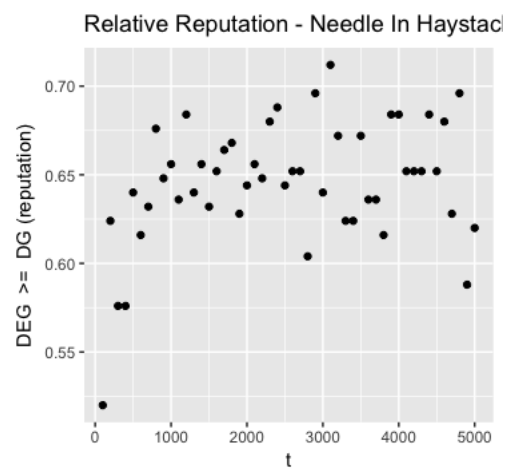
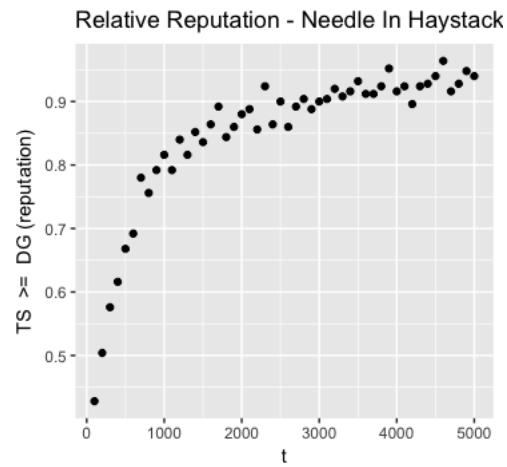
## Appendix

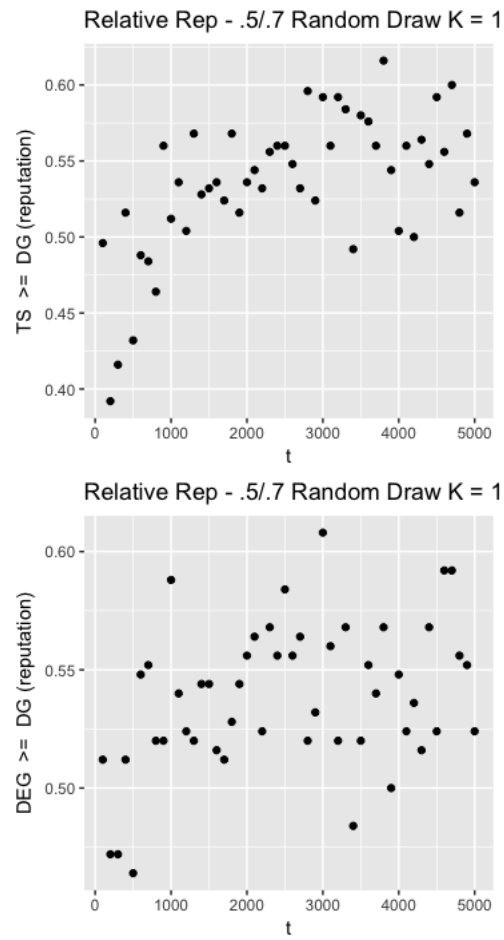
The rest of the relative reputation plots:











Finally, here is the warm start experiment redone with reputation erased after the warm start:

## Results for Heavy Tail HardMax K=10

	<b>WS = 5</b>	<b>WS = 20</b>	<b>WS = 50</b>	<b>WS = 100</b>	<b>WS = 200</b>	<b>WS = 400</b>	<b>WS = 700</b>	<b>WS = 1000</b>
TS vs DG	<b>0.34</b> (0.05) <u>eeog</u> avg: 91 med: 0	<b>0.41</b> (0.05) <u>eeog</u> avg: 250 med: 0	<b>0.55</b> (0.05) <u>eeog</u> avg: 470 med: 5	<b>0.67</b> (0.05) <u>eeog</u> avg: 580 med: 10	<b>0.71</b> (0.05) <u>eeog</u> avg: 700 med: 7	<b>0.73</b> (0.05) <u>eeog</u> avg: 830 med: 6	<b>0.76</b> (0.04) <u>eeog</u> avg: 890 med: 8.5	<b>0.77</b> (0.04) <u>eeog</u> avg: 860 med: 14
TS vs DEG	<b>0.3</b> (0.05) <u>eeog</u> avg: 28 med: 0	<b>0.43</b> (0.06) <u>eeog</u> avg: 200 med: 0	<b>0.64</b> (0.05) <u>eeog</u> avg: 360 med: 5	<b>0.82</b> (0.04) <u>eeog</u> avg: 400 med: 7	<b>0.83</b> (0.04) <u>eeog</u> avg: 470 med: 7.5	<b>0.83</b> (0.04) <u>eeog</u> avg: 580 med: 10.5	<b>0.85</b> (0.04) <u>eeog</u> avg: 720 med: 14	<b>0.85</b> (0.04) <u>eeog</u> avg: 660 med: 11
DG vs DEG	<b>0.57</b> (0.05) <u>eeog</u> avg: 910 med: 78.5	<b>0.59</b> (0.05) <u>eeog</u> avg: 870 med: 76.5	<b>0.62</b> (0.05) <u>eeog</u> avg: 1000 med: 89	<b>0.58</b> (0.05) <u>eeog</u> avg: 970 med: 88	<b>0.58</b> (0.05) <u>eeog</u> avg: 870 med: 25	<b>0.57</b> (0.05) <u>eeog</u> avg: 910 med: 25.5	<b>0.52</b> (0.05) <u>eeog</u> avg: 780 med: 16	<b>0.5</b> (0.05) <u>eeog</u> avg: 800 med: 26.5

[1] “

”

## Results for Uniform HardMax K=10

	<b>WS = 5</b>	<b>WS = 20</b>	<b>WS = 50</b>	<b>WS = 100</b>	<b>WS = 200</b>	<b>WS = 400</b>	<b>WS = 700</b>	<b>WS = 1000</b>
TS vs DG	<b>0.39</b> (0.05) <u>eeog</u> avg: 480 med: 11	<b>0.34</b> (0.05) <u>eeog</u> avg: 680 med: 12	<b>0.43</b> (0.05) <u>eeog</u> avg: 740 med: 11.5	<b>0.45</b> (0.05) <u>eeog</u> avg: 870 med: 28.5	<b>0.43</b> (0.05) <u>eeog</u> avg: 1100 med: 59.5	<b>0.56</b> (0.05) <u>eeog</u> avg: 1300 med: 308.5	<b>0.53</b> (0.05) <u>eeog</u> avg: 1200 med: 208.5	<b>0.53</b> (0.05) <u>eeog</u> avg: 1100 med: 128
TS vs DEG	<b>0.4</b> (0.05) <u>eeog</u> avg: 440 med: 7	<b>0.35</b> (0.05) <u>eeog</u> avg: 420 med: 6	<b>0.43</b> (0.05) <u>eeog</u> avg: 640 med: 14	<b>0.41</b> (0.05) <u>eeog</u> avg: 730 med: 15	<b>0.43</b> (0.05) <u>eeog</u> avg: 970 med: 21	<b>0.47</b> (0.05) <u>eeog</u> avg: 1200 med: 258.5	<b>0.51</b> (0.05) <u>eeog</u> avg: 1200 med: 337.5	<b>0.54</b> (0.05) <u>eeog</u> avg: 1300 med: 121
DG vs DEG	<b>0.51</b> (0.05) <u>eeog</u> avg: 1100 med: 102	<b>0.49</b> (0.05) <u>eeog</u> avg: 1400 med: 398	<b>0.45</b> (0.05) <u>eeog</u> avg: 1300 med: 378.5	<b>0.46</b> (0.05) <u>eeog</u> avg: 1400 med: 541.5	<b>0.53</b> (0.05) <u>eeog</u> avg: 1500 med: 795	<b>0.46</b> (0.05) <u>eeog</u> avg: 1500 med: 918.5	<b>0.46</b> (0.05) <u>eeog</u> avg: 1500 med: 706	<b>0.48</b> (0.05) <u>eeog</u> avg: 1300 med: 443

[1] “