1 Additional Experimental Results

1.1 Different number of total agents.

We set parameters: Time horizon $T = 4096$, Action set size $= 100$, dimension $d = 15$.

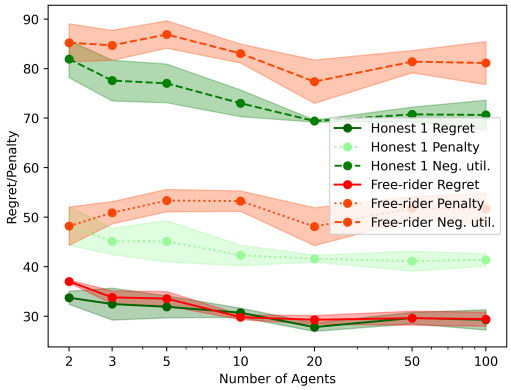


Figure 1: Penalties of a honest agent and free-rider as total number of agents vary.

We observe in Figure 1 that our mechanism consistently penalizes the free-rider for different number of agents $M > 2$.

For $M = 2$, the penalty is identical for the free-rider and the honest agent as, in our statistical test, the parameter estimate of each is compared against the other's and have symmetrical distances. In other words, with 1 liar and 1 truth-teller, the mechanism can't distinguish the lie as there is no numerical advantage of the truth.

1.2 Comparison with benchmark from literature.

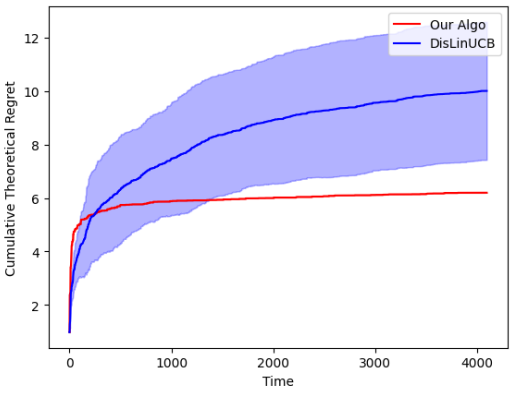


Figure 2: Regret comparison of our algo vs Distributed LinUCB algorithm Wang et al. [2019]

In Figure 2, we see our algorithm enjoys lesser regret than the baseline DisLinUCB [Wang et al., 2019] for collaborative linear bandits, it is to be noted that they use fewer communication rounds of data sharing.

While both works try to minimize regret, Wang et al. [2019] strive to limit the communication overhead, whereas we strive to be robust to strategic agents.

1.3 Different number of free-riders.

We set parameters: Time horizon $T = 2048$, Action set size $= 50$, dimension $d = 15$, total agents $M = 10$.

We observe that the ability to reliably penalize the free-rider slowly degrades as more free-riders enter the system. Notice the penalty curves in Row 2 of Figure 3 as we increase the number of free-riders, say f.

Initially, when f $= 1$, the free-rider gets a distinctly higher penalty than others. On increase to f $= 2$, the penalties of free-riders and honest agents get closer. With f $= 3, 4$, the penalties are no longer well separated between the honest agents and free-riders, who are a significant fraction of the populace here.

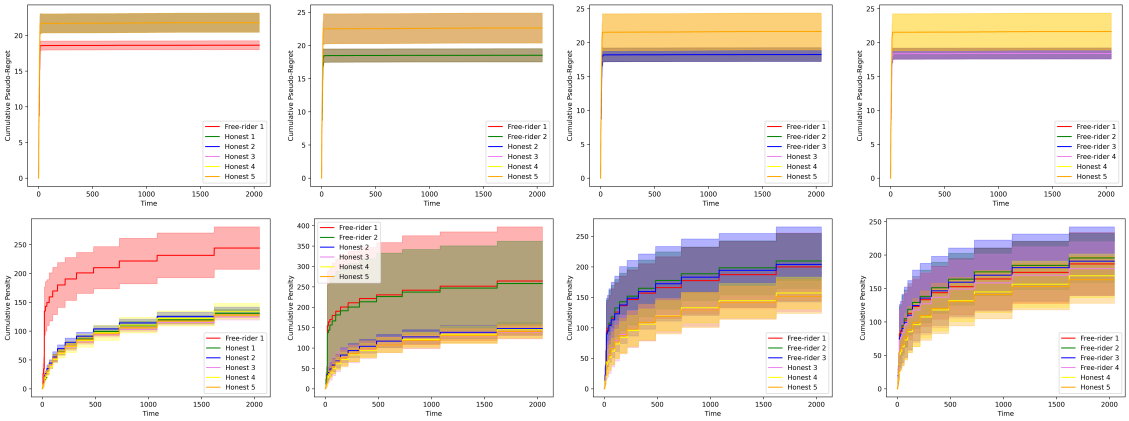
This can be attributed to the inaccuracy (or corruption) in the $\theta_a^{s,\varepsilon}$ parameter estimates that the principal protocol recovers and uses in the statistical tests. With more free-riders, more far-off are the design matrices used in computing these estimates, and the more unreliable are these estimates used in computing penalties.

This result is in line with our theoretical guarantee of Nash Equilibrium of all agents obeying the protocol—loosely, "it is good to be honest when all others are honest".

Other synthetic experiments reveal graphs and insights identical to the ones shared in the main paper submission.

References

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2019.



(a) f $= 1$ free-rider. (b) f $= 2$ free-riders. (c) f $= 3$ free-riders. (d) f $= 4$ free-riders

Figure 3: The regret (first row) and the levied penalties (second row) as the total number of free-riders, f, changes. The total number of agents is $M = 10$.