THE PERILS OF EXPLORATION UNDER COMPETITION: A COMPUTATIONAL MODELING APPROACH

GUY ARIDOR
COLUMBIA ECONOMICS

KEVIN LIU
COLUMBIA CS

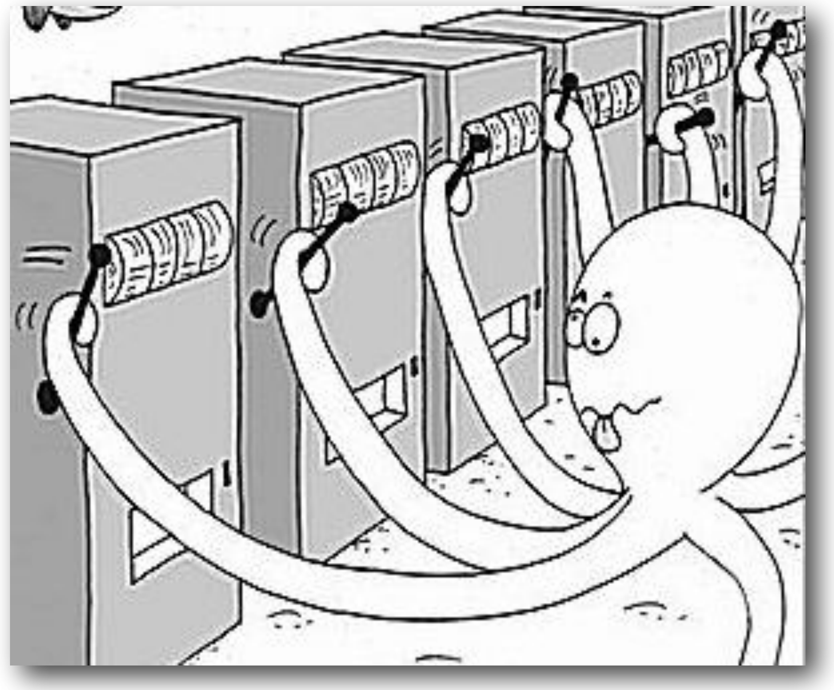
ALEX SLIVKINS
MICROSOFT RESEARCH, NYC

STEVEN WU
UNIVERSITY OF MINNESOTA CS

20TH ACM CONFERENCE ON ECONOMICS AND COMPUTATION

MOTIVATION

- Online platforms increasingly engage in ***product experimentation***
  - Search Engines
  - Recommender Systems
  - E-commerce platforms



- However, they also simultaneously ***compete for users***
- This paper: Firms **compete** for users and **learn** from the data generated by them

OUR SCOPE

- Study the tradeoff between ***exploration*** and ***competition.***
  1. Need to incentivize users to choose me over competition **today**
  2. Need to explore to gain information to have a better product **tomorrow**
- Questions:
  - Does competition incentivize adoption of better algorithms?
  - What is the role that data can play as barriers to entry?

(STOCHASTIC) MULTI-ARMED BANDITS

- In each period, select an action ("arm") from a fixed set of arms, observe (random) reward for this arm, and nothing else
  - mean reward of each arm is fixed over time but not known
  - Goal: maximize cumulative reward over **T** periods.
- Captures exploration-exploitation tradeoff
  - **Exploit -** Make the best decision today given the current information
  - **Explore** - Make a sub-optimal decision today (w.r.t. current information) in order to gather information and make *better decisions tomorrow*

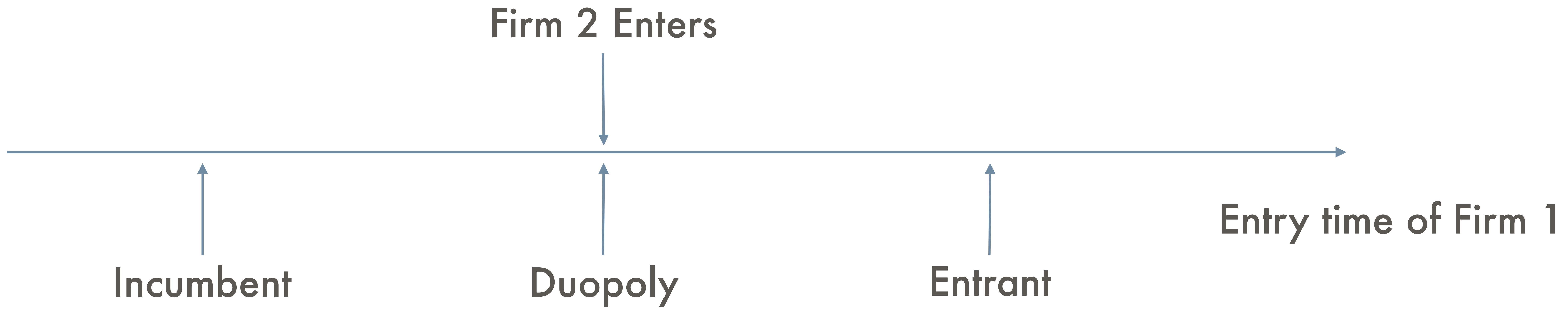
OUR MODEL

- **Two** firms, both face the same bandit problem
  - **K arms: different ways to serve a user**
  - **Initially, each firm commits to a bandit algorithm**
  - Warm start: T_0 rounds before the competition starts
- In each round: new user arrives and chooses a firm,
  the firm chooses an "arm", the user receives a reward
  - **Reward is only observed by the chosen firm**
- Each firm's goal: maximize its (**expected) market share**
- **User's choice driven by "reputation" (average reward over sliding window)**

INNOVATION VS COMPETITION

- **Innovation**: Utilize the distinction between three classes of **MAB learning algorithms.**
  - **Dynamic Greedy (DG):** pick arm with maximum mean reward based on current information.
  - **Exploration-Separating:** exploration does not use observations.
    - Use Dynamic Eps-Greedy (DEG): choose random arm with probability epsilon, else Greedy
  - **Adaptive Exploration:** zoom in on the best arm. Use Thompson Sampling (TS)

- **Competition**: vary **timing of entry** and **number of firms** in the market



METHODOLOGY

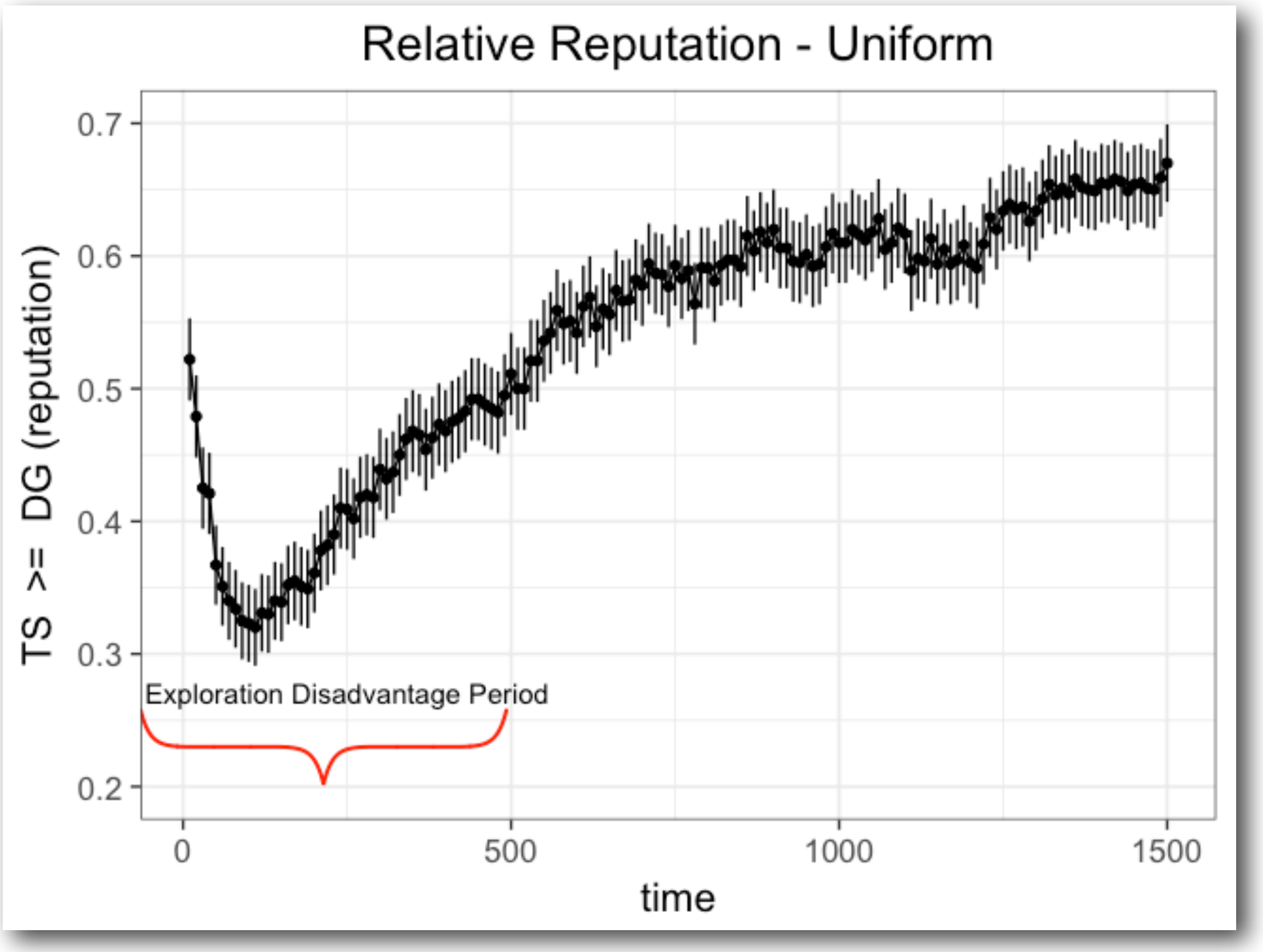
- Study our model via numerical simulation
- Consider three representative classes of instances:
  - Needle-In-Haystack - 1 “good” arm, K-1 identical “bad” arms
  - Uniform - mean rewards drawn from Uniform[0.25, 0.75]
  - Heavy Tail - mean rewards drawn from Beta(0.6, 0.6)
- <u>Each experiment</u>: competition between bandit algorithms
  - Parameters: bandit algorithms, competition model, bandit instance

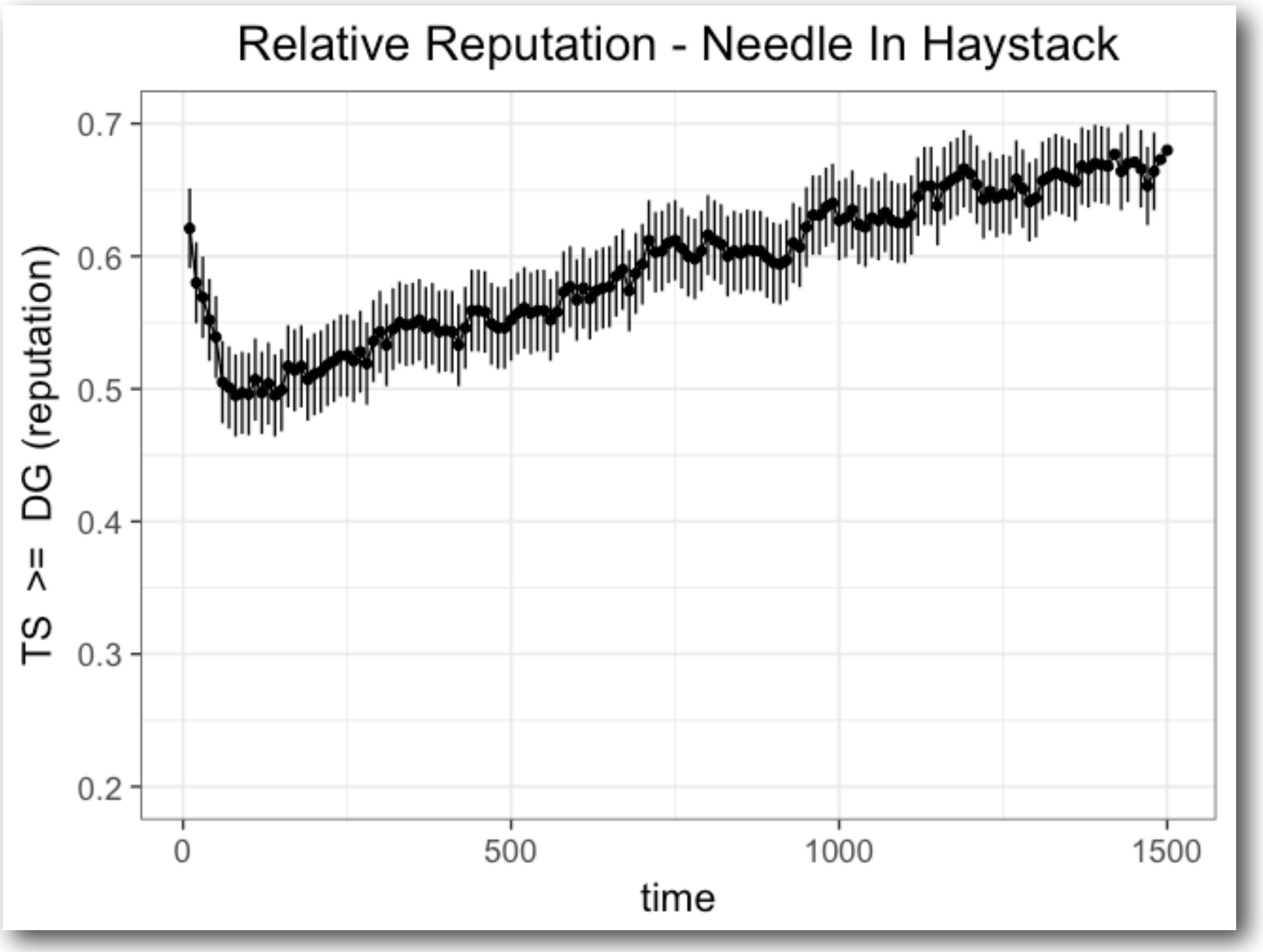
RELATED LITERATURE

- Multi-armed bandits: well-studied model for exploration
    - Huge literature on bandit algorithms
- Bandit algorithms with incentives (large literature, different scenarios):
    - "principal" runs a bandit algorithm
    - "agents" are bidders in an auction, users in a recommendation system, etc.
- Competition vs Innovation
    - In general: "Inverted-U" relationship: Schumpter (1942), Aghion et.al (2005)
    - For exploration: (Mansour, Slivkins, Wu 2018)
        - different model: no "reputation", competition varied via user response
        - Theory only, "asymptotic" results

PERFORMANCE IN ISOLATION

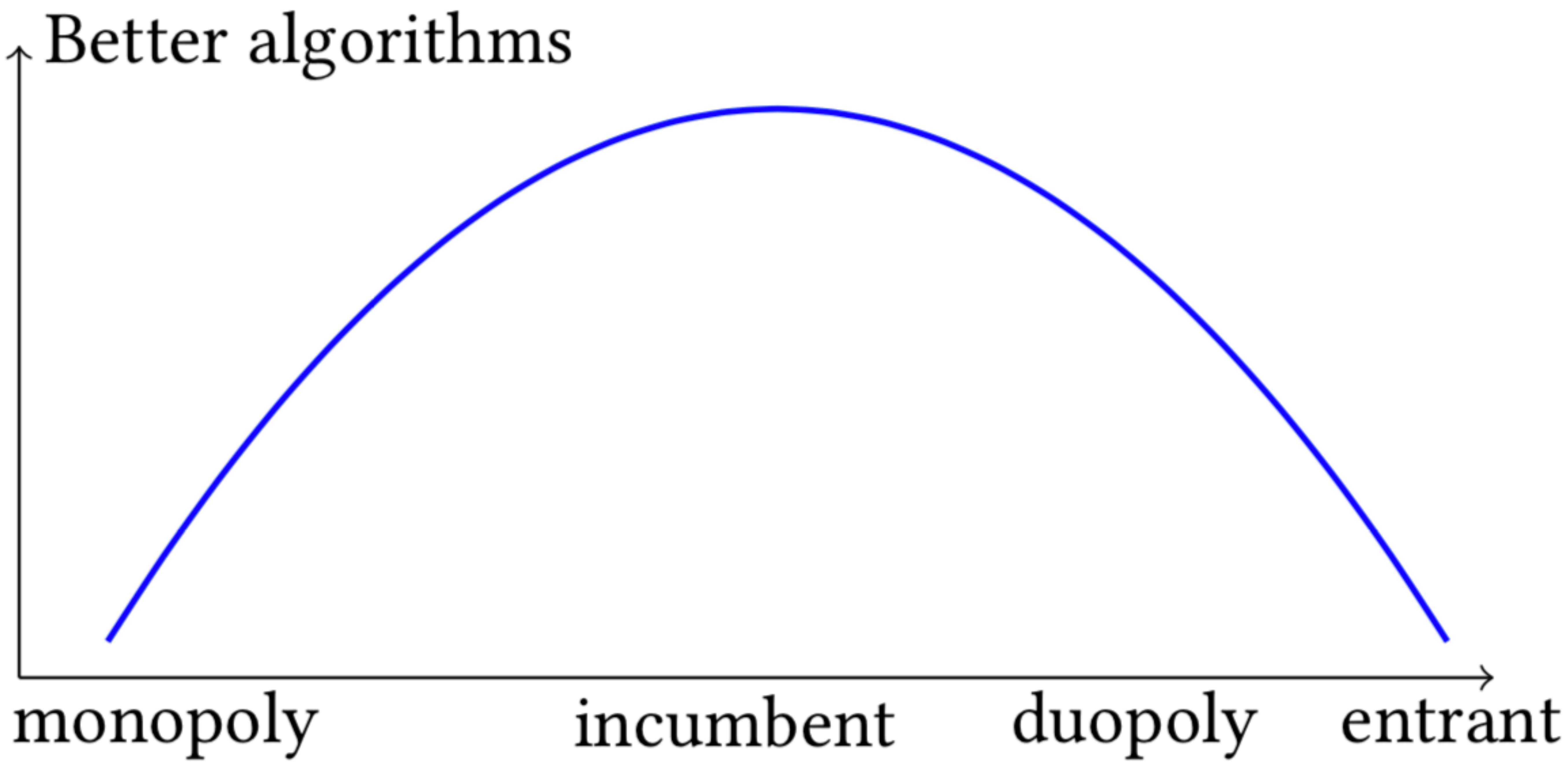
- Mean reputation not most predictive statistic for results in competition
- Better predictor: ***relative reputation*** - at a given time t, fraction of simulations in which Alg 1 has a higher reputation score than Alg 2
- Purposeful exploration can lead to short-term ***reputation consequences***
    - When this occurs, call the instance ***exploration-disadvantaged***





MAIN RESULTS

On exploration disadvantaged instances, we have the following set of results:



Stylized “inverted-U” relationship between competition and innovation

DUOPOLY (SIMULTANEOUS ENTRY)

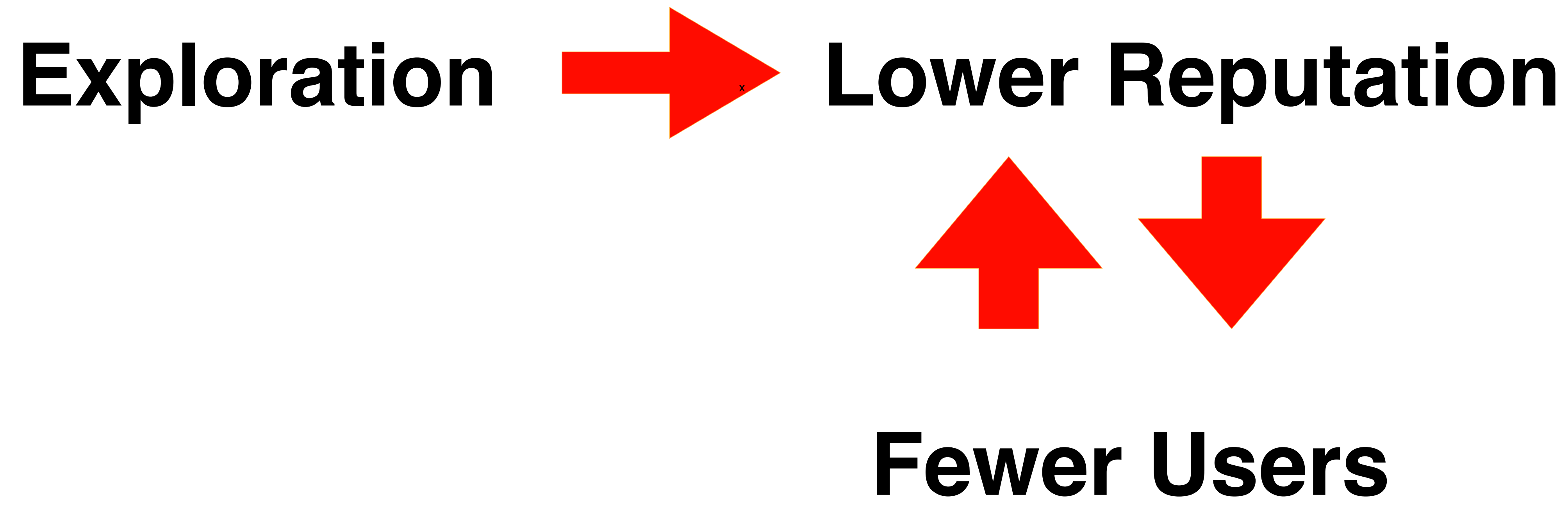
DEATH SPIRAL

- Greedy algorithms incentivized in equilibrium over "better" algorithms
- Effective End of Game: last round ***t*** s.t. agents ***t*** and ***t-1*** choose different firms

	TS vs DG	TS vs DEG	DG vs DEG
Effective End of Game	55 (0)	37 (0)	410 (7)

Mean (Median) Effective End of Game for Heavy-Tail Instance, $T = 2000$

- Low effective end of game indicative of ***death spiral effect:***



EARLY ENTRY

EQUILIBRIUM

- Allow one firm to enter early and give it a ***temporary monopoly***
  - ***Incumbent*** (the early entrant)***:*** Thompson Sampling is a dominant strategy
  - ***Entrant*** (the late entrant): Dynamic Greedy is a dominant strategy

	TS	DEG	DG
TS	$\mathbf{0.003}\pm0.003$	$\mathbf{0.083}\pm0.02$	$\mathbf{0.17}\pm0.02$
DEG	$\mathbf{0.045}\pm0.01$	$\mathbf{0.25}\pm0.02$	$\mathbf{0.23}\pm0.02$
DG	$\mathbf{0.12}\pm0.02$	$\mathbf{0.36}\pm0.03$	$\mathbf{0.3}\pm0.02$

User share of row player (entrant), 200 round head-start, Heavy-Tail Instance

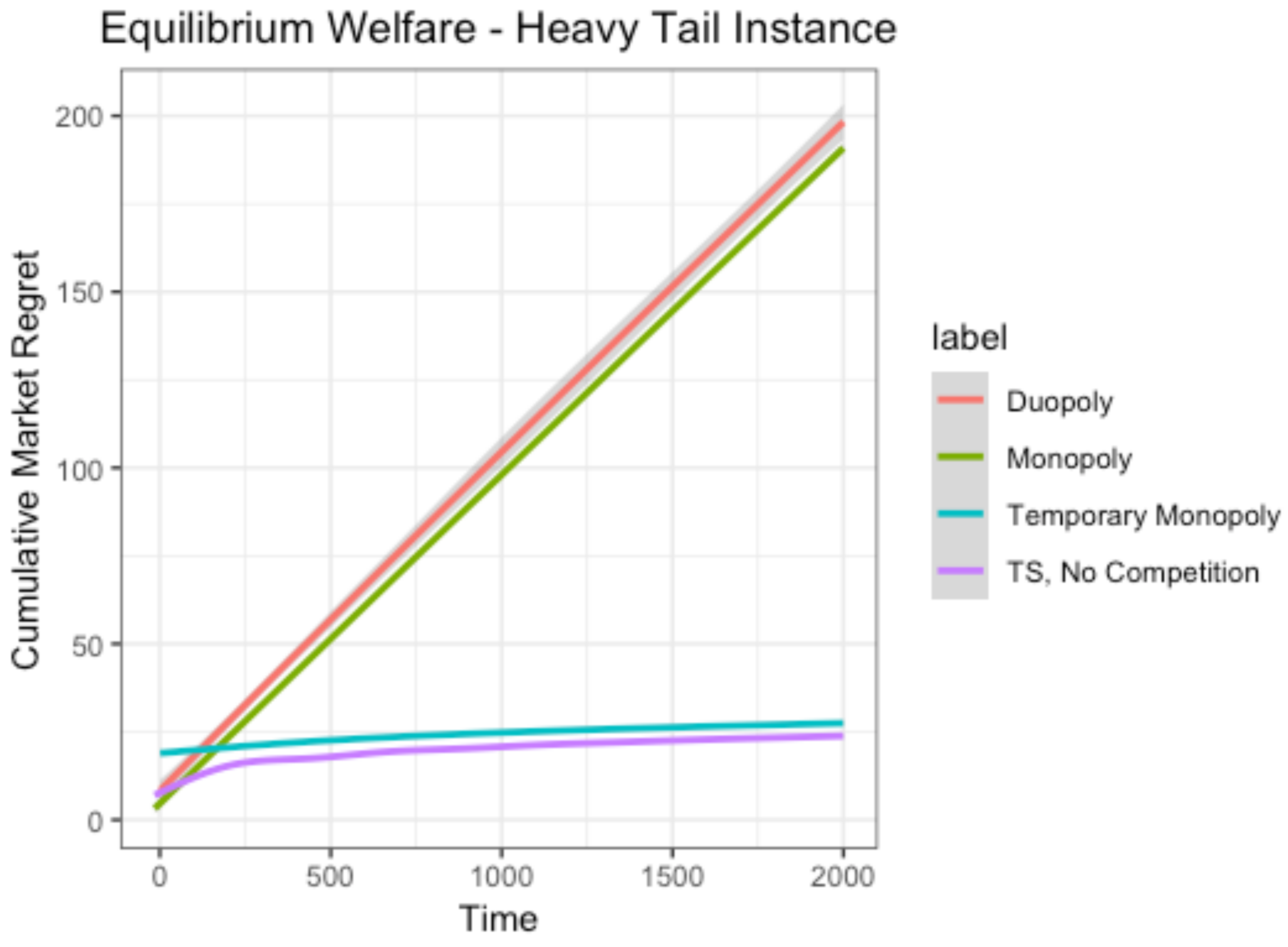
EARLY ENTRY

INTUITION

- Incumbent does not incur the immediate reputation consequences of exploration
- For sufficiently large "temporary monopoly" period,
  - incumbent only faces the classic exploration-exploitation tradeoff
  - picks algorithms that are best at optimizing this tradeoff
  - recovers the reputation consequences of exploration
  - still needs to compete against later entrant

WELFARE
EQUILIBRIUM

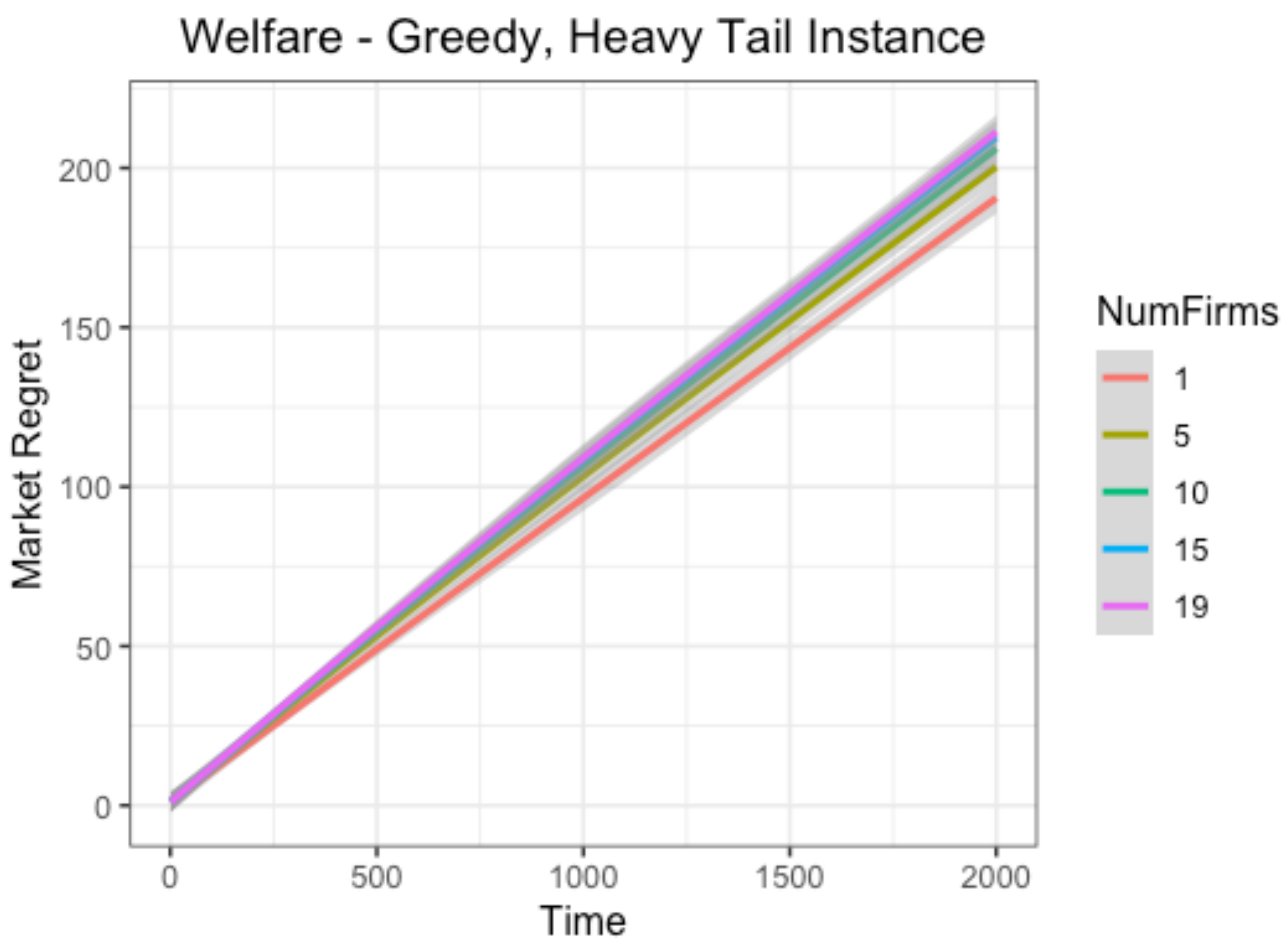
- Welfare measure = total “regret” accrued by all users
- Temporary monopoly induces highest welfare in competition



WELFARE

MANY FIRMS

- Restricting firms to playing greedy, increasing number of firms ***weakly decreases*** welfare



DATA AND REPUTATION AS BARRIERS TO ENTRY

- Two advantages of early entry:
  - ***Reputation advantage***: More definite / better reputation
  - ***Data advantage***: More data than the entrant
- Which advantage is a larger barrier to entry? Two experiments:
  - Reputation advantage only: reset incumbent's information when game starts
  - Data advantage only: reset incumbent's reputation when game starts

DATA OR REPUTATION?

- Either advantage alone leads to large market share
- Data advantage is larger when incumbent commits to Thompson Sampling

	Reputation advantage (only)			Data advantage (only)		
	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

User share of row player (entrant)

CONCLUSION

- Considered a model of competition between learning algorithms
- "Better algorithms" not always better in competition due to the reputational consequences of exploration
- Data can serve as a barrier to entry in online platforms, especially when exploration has reputational consequences