Incentivizing Exploration by Heterogeneous Users

COLT 2018

Bangrui Chen, Peter Frazier

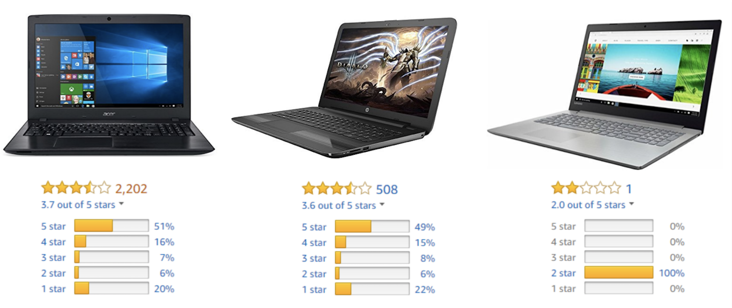
Cornell University
Operations Research and Information Engineering
bc496@cornell.com, pf98@cornell.edu

David Kempe

University of Southern California
Department of Computer Science
david.m.kempe@gmail.com

November 5, 2018

Customers Undervalue Exploration



- Incentives are misaligned:
  - Customers are myopic and want to **exploit**
  - Amazon wants customers to **explore**
- To fix this, Amazon can **incentivize exploration**

Previous Work

Without Money Transfer

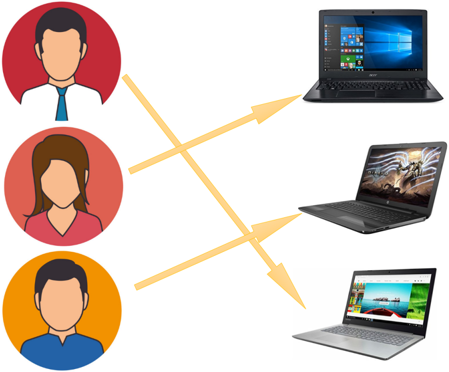
- Kremer, Mansour & Perry 2014
- Mansour, Slivkins & Syrgkanis 2015
- Mansour, Slivkins, Syrgkanis & Wu 2016
- Mansour, Slivkins & Wu 2018
- Slivkins 2017

With Money Transfer

- Frazier, Kempe, Kleinberg & Kleinberg 2014
- Han, Kempe & Qiang 2015

All of this work assumes agents have homogeneous preferences over items

We Incentivize **Heterogeneous** Agents



- **Our setting**: Customers have different preferences
- **Challenge**: Amazon doesn't know these preferences
- **Opportunity**: Heterogeneity provides free exploration

We Also Ask a Bigger Question

- Active exploration is critical in bandit theory
- Many practitioners don't do it
- **Why?**
- Do the imperfections of practice create free exploration?
- Other work: den Boer & Zwart 2015

Problem Setting

Agents

- Myopic agents arrive sequentially
- Agent t has linear utility with preference vector $\boldsymbol{\theta}_t \in \mathbb{R}^d$ drawn from known distribution F

Arms

- Each arm has an unknown feature vector $\boldsymbol{u}_i \in \mathbb{R}^d$
- Pulls give noisy observation of $\boldsymbol{u}_i$ with independent sub-Gaussian noise
- Everyone observes averages $\hat{\boldsymbol{u}}_{i,t}$ of each arm's past pulls

Problem Setting

Agents' behavior

- Principal chooses payment $c_{t,i}$ for arm i at time t
- Agent t pulls arm $i_t = \arg\max_i \{\boldsymbol{\theta}_t \cdot \hat{\boldsymbol{u}}_{i,t} + c_{t,i}\}$

Principal's Goal

- Regret: $r_t = (\max_i \boldsymbol{\theta}_t \cdot \boldsymbol{u}_i) - \boldsymbol{\theta}_t \cdot \boldsymbol{u}_{i_t}$
- Payment: $c_t = c_{t,i_t}$
- Minimize cumulative regret with small cumulative payment

Algorithm Sketch

An arm is **payment-eligible** if:

- without incentives, its probability of being pulled is below a threshold
- AND it hasn't been pulled in a long-time

Our **algorithm**:

- If there is a payment-eligible arm, offer enough incentive to raise its probability of being pulled above the threshold
- Otherwise, let agents play myopically

Algorithm Notation

- **Phase**: Phase s starts when each arm has been pulled at least s times.
- **Number of Pulls**: $m_{t,i}$ is the number of pulls of arm i up to time t.
- **Payment-eligible**: An arm i is *payment-eligible* at time t (in phase s) if:
  - i has been pulled at most s times up to time t, i.e., $m_{t,i} \leq s$.
  - AND the conditional probability of pulling arm i is less than $1/\log(s)$ given $\hat{u}_{t,i}$.

Our Algorithm:

Set the current phase number $s = 1$.
for time steps $t = 1, 2, 3, \ldots$ **do**
  if $m_{t,i} \geq s + 1$ for all arms i **then**
    Increment the phase $s = s + 1$.
  if there is a payment-eligible arm i **then**
    Let i be an arbitrary payment-eligible arm.
    Offer payment $c_{t,i} = \max_{\theta,i'} \theta \cdot (\hat{\mu}_{t,i'} - \hat{\mu}_{t,i})$ for pulling arm i (and payment 0 for all other arms).
  else
    Let agent play myopically, i.e., offer no payments.

Key Assumptions

- (**Every arm is someone's best**) Each arm is preferred by at least p fraction of users.
- (**Compact Support**) $\boldsymbol{\theta}$ has compact support.
- (**Few near-ties**) Let $q(z)$ be the proportion of agents with Utility(best arm) $\leq z +$ Utility(2^{nd} best arm).
  Then $q(z) \leq L \cdot z$ for all small enough z.

Main Result

Theorem 1

Our policy achieves:

- expected cumulative regret $O(Ne^{2/p} + LN\log^3(T))$,
- using expected cumulative payments of $O(N^2e^{2/p})$.

Discrete Preferences Give Constant Regret

Theorem 2

When agent preferences are discrete ($L = 0$), an algorithm using a modified algorithm has:

- expected cumulative regret $O(N^2/p)$,
- using expected cumulative payments of $O(N/p)$.

- Regret and payment are constant in T
- The classical MAB has regret $O(\log T)$
- **Heterogeneity gives free exploration**

Known p Gives Poly$(1/p)$ Regret/Payment

Theorem 3

When a lower bound on p is known, an algorithm using a modified threshold has:

- expected cumulative regret $O(\frac{N^2}{p^2} + \frac{NL\log^3(T)}{p})$,
- using expected cumulative payments of $O(N^2 \cdot \max(1, (L/p)^{5/2}))$.

Payment Analysis

Key technical lemma: An adaptive concentration inequality (Zhao et al. 2016).

Early Phases: Bound the number of payments in each phase by N.

Later Phases: Incentives are only needed when estimation error is large. Their probability shrinks exponentially as phases advance.

Regret Analysis

When principal incentivizes: similar to the payment proof

When agents pull myopically: We define a phase-dependent cutoff $\gamma(s(t))$ to separate agents with small and large regret.

- $r(t) \geq \gamma(s(t))$:
  - ⋆ this requires severe misestimates of arm attributes
  - ⋆ this happens with exponentially decreasing probability
  - ⋆ since θ_t has a compact support, the maximum regret is bounded by a constant
- $r(t) \leq \gamma(s(t))$:
  - ⋆ this requires nearly-tied preferences
  - ⋆ few agents have nearly-tied preferences
  - ⋆ the maximum regret is bounded above by $\gamma(s(t))$

Conclusion

- We provide the first incentivizing exploration analysis for agents with heterogeneous preferences over items
- When preferences are discrete and every arm is someone's best arm, regret and payments are constant in T
- Heterogeneity provides free exploration